Neural-Network Supervised Maximum Likelihood-based on-line Dereverberation

Saeed Mosayyebpour*, and Francesco Nesta*
*Synaptics, 1901 Main Street, Irvine, CA (USA). E-mail: {saeed.mosayyebpour,francesco.nesta }@synaptics.com

***Abstract*—In this paper, a new online multiple-input multiple-output (MIMO) approach based on Maximum Likelihood (ML) in subband-domain for dereverberation is proposed. Multi-channel linear prediction filters are estimated to blindly shorten the Room Impulse Responses (RIRs) between a set of unknown number of sources and a microphone array. The adaptive filter is updated using a modified weighted recursive Least Squares (RLS). To speed up convergence and minimize the influence of noise, the adaptive algorithm is supervised by a trained Deep Neural Network (DNN) which predicts the source dominance. In our experiments, it is proved that the proposed method can largely reduce the effect of reverberation in high non-stationary noisy conditions and sensibly improve automatic speech recognition performance in far-field and high reverberation.**

***Index Terms*—multiple-input multiple-output (MIMO); Maximum Likelihood (ML); dereverberation; recursive Least Squares (RLS) ; Deep Neural Network (DNN);**

I. Introduction

There are many offline reverberation reduction solutions available in the literature (e.g. [1]-[7]). However, many of them can not be used in real-time applications as they require long buffer of data to compensate the effect of reverberation or to estimate inverse filters [1]. In addition, some methods require a high amount of memory and are not computationally efficient for low power devices used in embedded applications.

In order to make dereverberation possible in many practical industrial applications, a number of online adaptive algorithms have been developed (e.g. [8]-[10]). In [8]-[9] the authors employed Recursive Least Squares (RLS) method to develop the adaptive Weighted Prediction Error (WPE) approach. A Kalman filter approach is proposed in [10] where a multi-microphone algorithm simultaneously estimates the clean speech signal and the time-varying acoustic system. In [10], the recursive expectation-maximization scheme is employed to obtain both the clean speech signal and the acoustic system in an online manner. Despite the proposed algorithms have been shown to effectively reduce the reverberation in an on-line fashion, both methods do not explicitly model the noise and can underperform in highly non-stationary noisy conditions. In addition, the computational complexity and memory usage for the Kalman algorithm might be unreasonably too high to be implemented in a low cost embedded device.

In order to fill the gaps and produce a robust on-line solution for real-world dereverberation, in this work an ML-based adaptive algorithm using the RLS method is proposed. To estimate the prediction filters in an online-manner, a new weighted cost function is proposed which is minimized at each frame when the speech source is active. The estimation is supervised using a binary speech presence posterior weight which is produced by a trained Deep Neural Network (DNN). In order to improve the accuracy and the convergence speed of the on-line algorithm, a deterministic function is used to model the power decay of the late reverberation and the presence of additive background noise is explicitly accounted in the model. The algorithm produces a linearly filtered multichannel speech signal which is non-linearly post-processed and then fed to a multichannel noise suppression based on the S-IVA approach [11], in order to further reduce the presence of noise and improve the overall performance.

The proposed method has the following advantages over other online-based dereverberation methods.

- It is more robust to reverberation and noise as the proposed online method reduces the effect of reverberation in two steps namely linear filtering and nonlinear filtering similar to what is proposed in [6] for batch processing.
- It has better and more accurate estimation of dereverberation filter in high non-stationary noisy condition by taking the advantage of a target source dominance estimation, which is obtained by training a neural network using extensive prior acoustic data.
- It has fast convergence rate through the use of modified RLS algorithm.
- It is a MIMO algorithm with no latency and therefore it can be easily integrated to other linear multichannel noise reduction methods.

Experimental evaluations with real-world data shows that the proposed method outperform other existing algorithms when using standard objective metrics for dereverberation performance. Furthermore, it is shown a considerable improvement of word recognition in a standard industrial test for far-field automatic speech recognition, when tested in presence of high reverberation and non-stationary noise.

II. Problem formulation

Let's assume the input signal for i-th channel is denoted by $x_i[n]$ $(i = 1, 2, ..., M)$ where M is the number of microphones and it is assumed that there is one source. Then the input signal can be modeled in frequency domain as ([4])

$$X_i(l,k) = \sum_{\acute{l}=0}^{L-1} H_i(\acute{l},k)S(l-\acute{l},k) + \nu_i(l,k)\ i = 1, 2, ..., M,$$

$$= \sum_{\acute{l}=0}^{D-1} H_i(\acute{l},k)S(l-\acute{l},k) + \sum_{\acute{l}=D}^{L-1} H_i(\acute{l},k)S(l-\acute{l},k) + \nu_i(l,k),$$

$$= Y_i(l,k) + R_i(l,k) + \nu_i(l,k), \quad (1)$$

ISBN 978-90-827970-1-5 © EURASIP 2018

where $S(l,k)$, $\nu(l,k)$ and $H_i(l,k)$ are clean speech signal, background noise signal and the RIR between the source and i-th microphone in frequency domain, respectively. L, l, and k are the length of the RIR, the frame index, and the frequency-bin index, respectively. In (1), the RIR is separated into two parts namely early reverberation which has length D and late reverberation part. So the corresponding signal is named as the desired speech and is denoted by $Y_i(l,k)$ and the second term is named as the reverberant speech and is denoted by $R_i(l,k)$. The goal is to extract the first term by reducing the second and the third term in noisy conditions.

In (1), both the clean speech and RIR are unknown. To simplify the task of the dereverberation, $R_i(l,k)$ is approximated by using the multichannel autoregressive model as given below [4]

$$R_i(l,k) \approx \sum_{\acute{l}=D}^{L+D-1} \mathbf{W}_i(\acute{l}-D,k)\mathbf{X}(l-\acute{l},k) + R_i^{rev}(l,k), \quad (2)$$

where $R_i^{rev}(l,k)$ is the residual late reverberation that cannot be reduced by linear filtering [6]. In (2) the unknown parameter to be estimated is an $M \times 1$ prediction filter vector ($\mathbf{W}_i(l,k) = [W_{i1}(l,k), ..., W_{iM}(l,k)]^T$) and $\mathbf{X}_i(l,k) = [X_1(l,k), ..., X_M(l,k)]^T$.

To estimate the prediction filter, the Maximum Likelihood (ML) approach is used based on three important assumptions [3],[4].

- The received speech signal has a Gaussian Probability Density Function (pdf) and the clean part of the received speech and noise has zero mean with time-varying variance.
- The frames of the input signal are independent random variables.
- The RIR is static or it changes slowly.

The ML cost function $L(.)$ using the aforementioned assumptions for a given received speech signal of T frames that is denoted by $\tilde{\mathbf{X}}(k)$ is given below

$$L(\tilde{\mathbf{X}}(k)|\mathbf{W}(l,k)) = \quad (3)$$
$$-\sum_{l=0}^{T-1} \{log(|\Sigma(l,k)|) + (\mathbf{X}(l,k) - \boldsymbol{\mu}(l,k))^H \Sigma(l,k)^{-1} (\mathbf{X}(l,k) - \boldsymbol{\mu}(l,k))\}$$

where $\Sigma(l,k)$ is the $M \times M$ spatial correlation matrix. Also, according to the above three assumptions, the mean $\boldsymbol{\mu}(l,k)$ can be approximately obtained as

$$\boldsymbol{\mu}(l,k) = [\mu_1(l,k), ..., \mu_M(l,k)]^T \quad (4)$$

$$\mu_i(l,k) = \sum_{\acute{l}=D}^{L+D-1} \mathbf{W}_i(\acute{l}-D,k)^H \mathbf{X}(l-\acute{l},k) \quad (5)$$
$$= \overline{\mathbf{W}}_i(k)^H \overline{\mathbf{X}}(l,k) \quad (6)$$

$$\overline{\mathbf{X}}(l,k) = [X_1(l-D,k), ..., X_1(l-D-L+1,k), ...$$
$$, X_M(l-D,k), ..., X_M(l-D-L+1,k)]^T \quad (7)$$

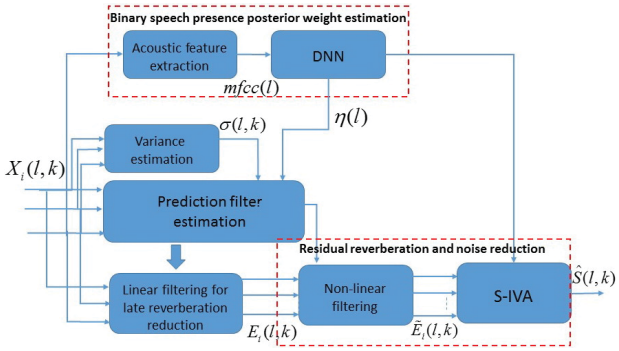

Fig. 1. The block diagram of the proposed method.

$$\overline{\mathbf{W}}_i(k) = [W_1^i(0,k), ..., W_1^i(L-1,k), ..., W_M^i(0,k), ..., W_M^i(L-1,k)]^T \quad (8)$$

where $\overline{\mathbf{W}}_i(k)$ is the prediction filter vector for frequency band k and i-th channel.

In order to estimate the prediction filter in an online manner, $\Sigma(l,k)$ is approximated by scaled identity matrix with the scale variance of $\sigma(l,k)$ (i.e. $\Sigma(l,k) = \sigma(l,k)I_{M\times M}$). With this assumption, (3) and (4) can be simplified as a weighted Mean Square Error (MSE) optimization problem and the simplified offline cost function $C_i(k)$ for the i-th channel can be written as

$$C_i(k) = \sum_{l=0}^{T-1} \frac{(X_i(l,k) - \mu_i(l,k))^2}{\sigma(l,k)}. \quad (9)$$

To estimate the prediction filters in an online-manner, the weighted MSE cost function will be minimized by updating $\mathbf{W}_i(k)$ at each frame when the source is active. In the absence of speech, the mean $\boldsymbol{\mu}(l,k)$ in (4) is equal to zero and so there is no need to update the prediction filters. This can be obtained by multiplying the cost function by a binary speech presence posterior weight $\overline{\eta}(l)$. Furthermore, in order to achieve faster convergence, the cost function is revised by using a forgetting factor ($0 < \lambda < 1$). Therefore the revised offline cost function can be written as

$$C_i(k) = \sum_{l=0}^{T-1} \overline{\eta}(l)\lambda^{T-l} \frac{(X_i(l,k) - \mu_i(l,k))^2}{\sigma(l,k)}. \quad (10)$$

III. Proposed Method

The input signals $x_i[n]$ are first passed through the sub-band decomposition to obtain the frequency domain signals $X_i(l,k)$. In this paper, the frame size of 25 ms with 4 ms frame shift is used to transform the signal from time-domain to the frequency domain. The block diagram of the proposed method is shown in Fig. 1. In the following sections, more detail is given for each stage of the processing.

A. Binary speech presence posterior weight estimation

A neural network (NN) is used to predict the source activity posteriors $\eta(l)$. Namely, the network is trained to estimate the power ratio between the true target speech and the noisy mixture. Any machine learning method can be used, such as

ISBN 978-90-827970-1-5 © EURASIP 2018

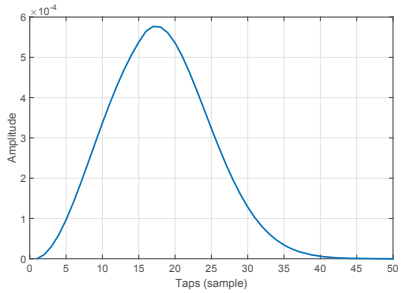

Fig. 2. Late reverberation weights in (15) when $b = 4$, $\acute{L} = 35$, $L = 45$ and $\rho = 0.01$

recurrent neural networks but we found that a naive multilayer feed forward NN, often named DNN, is sufficiently accurate to produce a useful prediction.

As it is clear from Fig. 1, the acoustic feature is estimated using the subband signals. In this paper, the MFCC feature plus the delta and double delta each having 12 coefficients are used to obtain the acoustic feature for training the DNN. Also the MFCC features of three consecutive non-overlapping frames are combined to form the input feature $F(l)$ for the DNN. Two hidden layers of 256 neurons are used with the hyperbolic tangent as the activation function. The softmax function is used in the output layer, which has dimension 2. The relative energy of clean versus the noisy signal is used as labels for training the DNN.

In this work, we focus on the scenario where the source of interest is "speech" while any other non-speech acoustic event is considered as "noise". For the training of the DNN, a large set of 100k mixtures was generated by randomly combining noise examples with speech sentences in the TIMIT database. Noises were collected from different sources and the dataset was designed to balance the amount of noises belonging to different categories. Noise signals selected did not contain any speech, as the scope of the network is only to discriminate between speech and noise. Two datasets of 10k mixtures were generated for both cross-validation and testing. After training, the first channel is used to predict speech presence posterior weight at the lth frame, indicated as $\eta(l)$ which is obtained through the feed-forward propagation of the input features. Finally the binary speech presence posterior weight is obtained as $\overline{\eta}(l) = (\eta(l) > \alpha)$, where α is a threshold with values between 0 and 1 (e.g. 0.5).

B. Input variance estimation

As it is clear from (1), the input variance $\sigma(l,k)$ has three components related to desired speech $Y_i(l,k)$ ($\sigma_y(l,k)$), the reverberant speech $R_i(l,k)$ ($\sigma_r(l,k)$) and the noise signal $\nu_i(l,k)$ ($\sigma_\nu(l,k)$). Similar to (26) in [6] and by using the assumption of mutually uncorrelated signals for each component of (1), the input variance can be approximated as

$$\sigma(l,k) \approx \sigma_y(l,k) + \sigma_r(l,k) + \sigma_\nu(l,k). \tag{11}$$

Below we describe how to estimate each component in an efficient way.

1) The first part is the variance for the early reflections. By using the current estimate for the prediction filter $W_m^i(l,k)$, the linearly dereverberated output signal $E_i(l,k)$ (i.e. the early reflection signal) is computed as

$$E_i(l,k) = X_i(l,k) - \sum_{m=1}^{M} \sum_{\acute{l}=0}^{L-1} X_m(l-D-\acute{l},k) W_m^{i\,*}(\acute{l},k). \tag{12}$$

The variance is then computed as the average over all the channels

$$\sigma_y(l,k) = \frac{1}{M} \sum_{i=1}^{M} E_i(l,k). \tag{13}$$

2) The second part is the variance of the late reverberation. This variance can be estimated using the following equation

$$\sigma_r(l,k) = \frac{1}{M} \sum_{\acute{l}=0}^{L-1} \widetilde{W}_l(\acute{l},k) \sum_{m=0}^{M-1} |X_m(l-D-\acute{l},k)|^2, \tag{14}$$

where $\widetilde{W}_l(\acute{l},k)$ is the late reverberation weights for l-th frame which is an unknown parameter. To speed up the on-line convergence, a fixed Rayleigh decay function, similar to the work presented in [2], is used to model the weights

$$\begin{aligned} R(\acute{l}) &= \frac{\acute{l}}{b^2} e^{\frac{-\acute{l}}{2b^2}}, \quad \acute{l} = 0, ..., \acute{L} \\ R(\acute{l}) &= 0, \quad \acute{l} = \acute{L}+1, ..., L \\ \widetilde{W}(\acute{l},k) &= \frac{\rho}{L-\acute{L}} \sum_{j=0}^{L-\acute{L}-1} R(\acute{l}-j), \end{aligned} \tag{15}$$

where $b = 4$, $\acute{L}_k = 35$ and $\rho = 0.01$ are the Rayleigh function parameter, the Rayleigh function length and the residual reverberation factor, respectively. Fig. 2 shows an example of the late reverberation weights.

3) The noise variance $\sigma_\nu(l,k)$ is estimated as a recursive smoothed input power spectrum averaged over all the channels, when the target speech is not active, i.e. when $\overline{\eta} = 0$.

C. Prediction filter estimation

To estimate the prediction filter for each channel recursively, the cost function in (10) is minimized. The update rule is given below.

$$w_m(0,k) = 0 \qquad \Phi(0,k) = \gamma I_M \tag{16}$$

$$RLS_{gain}(l,k) = \frac{\overline{\eta}(l)\Phi(l-1,k)\overline{\mathbf{X}}(l,k)}{\lambda\sigma(l,k) + \overline{\eta}(l)\overline{\mathbf{X}}^H(l,k)\Phi(l-1,k)\overline{\mathbf{X}}(l,k)} \tag{17}$$

$$\mathbf{W}_i^{(l)} = \mathbf{W}_i^{(l-1)} + RLS_{gain}(l,k) E_i^*(l,k) \tag{18}$$

$$\Phi(l,k) = \frac{\Phi(l-1,k) - RLS_{gain}(l,k)\overline{\mathbf{X}}^H(l,k)\Phi(l-1,k)}{\lambda} \tag{19}$$

where γ is a regularization factor that is set to 0.01. Here the delay and the forget factor are set to $D = 2$ and $\lambda = 0.998$, respectively. Also the prediction filter length L is set to 45.

ISBN 978-90-827970-1-5 © EURASIP 2018

Method	SNR=20 dB 0.5 m	1 m	2 m	3 m	4 m	avg	SNR=5 dB 0.5 m	1 m	2 m	3 m	4 m	avg
unproc	87.60	83.41	73.75	64.57	68.92	75.65	34.14	23.19	13.37	18.68	10.31	19.93
denoise	90.66	89.53	79.39	77.62	72.14	81.86	52.50	41.71	24.80	22.06	18.20	31.85
prop-noVAD	91.30	91.63	86.96	87.28	84.38	88.31	67.79	68.12	53.30	46.22	47.50	56.58
prop	91.95	91.30	86.80	86.63	84.82	88.30	69.89	68.44	54.59	47.83	48.79	57.90

TABLE I

THE WORD ACCURACY SCORES IN PERCENTAGE OBTAINED IN DIFFERENT NOISY REVERBERATION CONDITION FOR THE UNPROCESSED SIGNAL "UNPROC", PROPOSED METHOD "PROP", THE PROPOSED METHOD WHEN THE SPEECH PRESENCE POSTERIOR IS NOT USED "PROP-NOVAD", AND THE S-IVA METHOD ONLY "DENOISE".

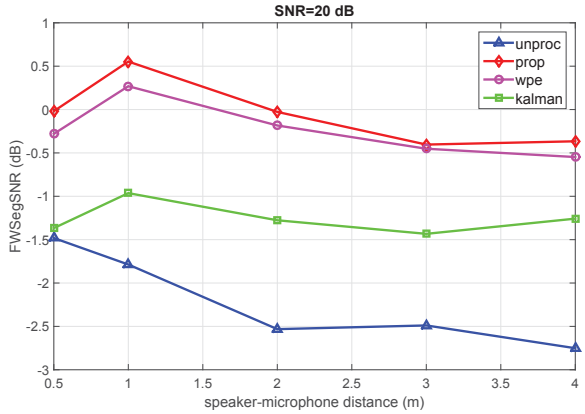

Fig. 3. The FWSegSNR performance of unprocessed signal "unproc", proposed method "prop", method of [8]-[9] (wpe) and [10] "kalman" in different noisy reverberant conditions when SNR=20 dB.

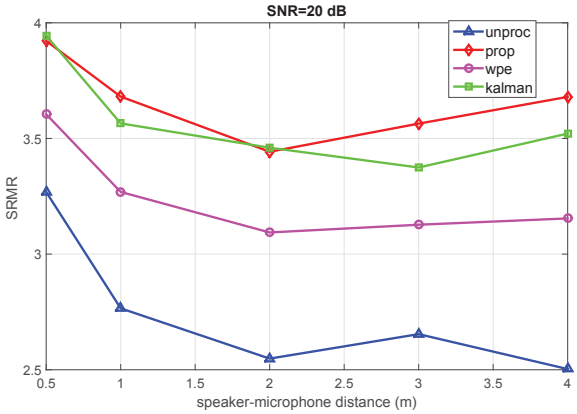

Fig. 5. The SRMR performance of unprocessed signal "unproc", proposed method "prop", method of [8]-[9] "wpe" and [10] "kalman" in different noisy reverberant conditions when SNR=20 dB.

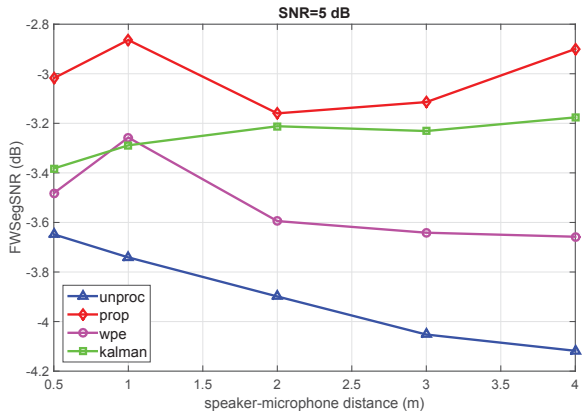

Fig. 4. The FWSegSNR performance of unprocessed signal "unproc" proposed method "prop", method of [8]-[9] "wpe" and [10] "kalman" in different noisy reverberant conditions when SNR=5 dB.

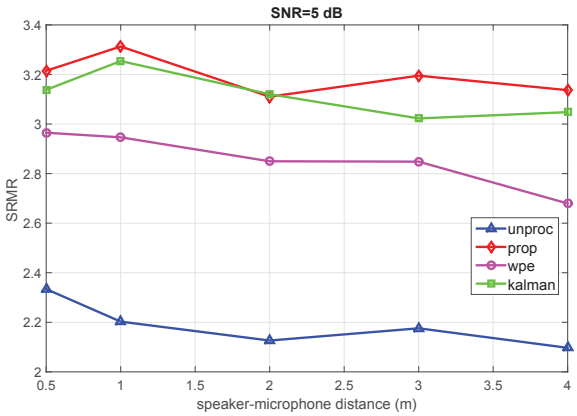

Fig. 6. The SRMR performance of unprocessed signal "unproc", proposed method "prop", method of [8]-[9] "wpe" and [10] "kalman" in different noisy reverberant conditions when SNR=5 dB.

D. Residual reverberation and noise reduction

Similar to [6], to reduce the residual late reverberation $R_i^{rev}(l,k)$ in (2) and improve the performance of reverberation reduction, a heuristic spectral gain $g(l,k) = \frac{(\sigma_y(l,k)+\sigma_\nu(l,k))}{\sigma(l,k)}$ is applied to the linear filtered signal $E_i(l,k)$ to further reduce spectrally the residual late reverberation as

$$\widetilde{E_i}(l,k) = E_i(l,k)g(l,k). \tag{20}$$

From the derverberated signals, the background noise is reduced by using the S-IVA algorithm described in [11], which produces the enhanced speech signal $\widehat{S}(l,k)$. Finally, the enhanced speech signal is transformed back to time-domain using the subband synthesis.

IV. EXPERIMENTAL EVALUATION

In this section, the performance of the proposed method in different noisy reverberant environments is evaluated. The results are obtained using about 8 min of clean speech signal which is not included in the training set of the DNN model. This clean signal is played by a loudspeaker and the signals are recorded at fs = 16 kHz in a room of size $5 \times 5 \times 2.5$ m with RT60 of about 430 ms. The recorded reverberant speech

ISBN 978-90-827970-1-5 © EURASIP 2018

signals are obtained by two microphones with mutual distance of 0.08 m when the clean speech is played by a loudspeaker at different distances of 0.5 m, 1 m, 2 m, 3 m and 4 m from the center of the microphones. Noisy mixtures are obtained by playing a stereo TV noise (not included in the training set of the DNN model) in order to produce a noisy signal at 20 dB and 5 dB SNRs. The thresholds α to obtain the binary speech presence posterior was set to the values of 0.7 and 0.9, when used with the proposed RLS and S-IVA algorithms, respectively.

To evaluate the propose method, two objective measures are utilized namely Frequency-weighted segmental SNR (FWSegSNR) [12] and Speech-to-reverberation modulation energy ratio (SRMR) [13]. FWSegSNR is based on the discrepancy between target and reference signals and it is obtained using the critical band analysis with the mel-frequency filterbank [12]. The FWSegSNR measure is highly correlated with the perceptual speech quality [12]. SRMR is a non-intrusive metric for speech quality and intelligibility based on a modulation spectral representation of the speech signal [13]. The results are averaged over the speech frames and larger values for both measures indicate better speech quality. In this paper, the proposed method "prop" is compared with methods presented in [8]-[9] "wpe" and [10] "kalman" and with the unprocessed signal "unproc". To make the comparison fair for the competing approaches, the noise reduction method (S-IVA) along with the pilot signal obtained through the DNN posterior is applied to the processed signal of "wpe" and "kalman" methods. The results are shown in Fig. 3-6. From Fig. 3-4, it is clear that the "wpe" method has better FWSegSNR performance than the "kalman" method in high SNR but it becomes worse in low SNR conditions. When the SRMR is considered (Fig. 5-6), the "kalman" method outperforms the "wpe" in both high and low SNR. On the other hand, the proposed method consistently outperforms both the "wpe" and the "kalman" method in both high and low SNR conditions and for any measurement, which highlights its robustness.

Finally the performance of proposed method is evaluated using a standard industrial ASR test, namely the Microsoft Cortana 1.0 score test. This is a benchmark originally designed for evaluating ASR performance of pc/laptops in near-field conditions and therefore is expected to show degradation with a high reverberation [1]. To highlight the effectiveness of the proposed dereverberation algorithm and also to show the effect of using the binary speech presence posterior, the proposed method "prop" is now compared with the noise reduction S-IVA method alone "denoise" [11] and the proposed method when the DNN posteriors are not used to control the update of the prediction filters "prop-noVAD". The absolute word accuracy rate (WAR) in percentage is shown in Table 1 where the average scores "avg" over all the 6 different distances for each SNR are also shown to facilitate the comparison. Comparing the average scores for "denoise" and "prop" indicates that the proposed dereverberation method could improve the ASR scores in both SNR conditions. Also, comparing the average scores for "prop" and "prop-noVAD" shows that the binary speech presence posterior for prediction filter estimation can sensibly improve the performance especially in low SNR. This is expected since in low SNR the prediction filter estimation can be largely affected by loud background noise.

[1]Newer versions of the test using far-field ASR models are available but were not considered in this work as it was not compatible with the data available at the time of writing this manuscript.

V. Conclusions

A new multichannel dereverberation algorithm based on the Maximum Likelihood estimation is proposed, in order to reduce the late the reverberation in noisy environments. The problem is formulated for an on-line solution using a modified adaptive RLS algorithm where the update of the filters is controlled by speech activity posteriors, estimated through a pretrained DNN. Unlike other counterparts, the proposed algorithm is robust to low SNR conditions as the noise is explicitly modeled in the cost function. Experimental results shows that the proposed method can consistently improve objective dereverberation benchmarks and drastically increase ASR scores when compared with other existing state-of-art methods for on-line dereverberation.

References

[1] S. Mosayyebpour, H. Sheikhzadeh, T. A. Gulliver, and M. Esmaeili, "Single-microphone LP residual skewness-based approach for inverse filtering of room impulse response,"*IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1617–1632, July 2012.

[2] S. Mosayyebpour, M. Esmaeili, and A. Gulliver, "Single- Microphone Early and Late Reverberation Suppression in Noisy Speech,"*IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 322–335, Feb. 2013.

[3] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction,"*IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1717–1731, Sep. 2010.

[4] Takuya Yoshioka and Tomohiro Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening,"*IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.

[5] Takuya Yoshioka, Tomohiro Nakatani, and Masato Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation,"*IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 231–246, Feb. 2009.

[6] M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga, "Optimized speech dereverberation from probabilistic perspective for time varying acoustic transfer function,"*IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1369–1380, July 2013.

[7] A. Jukic, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors,"*IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1509–1520, July 2015.

[8] T. Yoshioka, H. Tachibana, T. Nakatani, M. Miyoshi, "Adaptive dereverberation of speech signals with speaker-position change detection," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 394–406, Apr. 2009.

[9] Takuya Yoshioka, Tomohiro Nakatani, "Dereverberation for reverberation-robust microphone arrays," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Apr. 2013.

[10] B. Schwartz, S. Gannot, and E.A.P. Habets, "Online Speech Dereverberation Using Kalman Filter and EM Algorithm,"*IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 2, pp. 394–406, Feb. 2015.

[11] Francesco Nesta, Saeed Mosayyebpour, Zbynek Koldovsk, and Karel Palecek, "Audio/video supervised independent vector analysis through multimodal pilot dependent components," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Apr. 2017.

[12] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement,"*IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[13] Tiago H. Falk ; Chenxi Zheng ; Wai-Yip Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech,"*IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, Jan. 2010.

ISBN 978-90-827970-1-5 © EURASIP 2018