ADAPTIVE HIDDEN MARKOV MODELS FOR NOISE MODELLING

Jiongjun Bai and Mike Brookes

EEE Dept, Imperial College London
Exhibition Road, London SW7 2BT, UK
email: {jiongjun.bai04, mike.brookes}@imperial.ac.uk
web: http://www.commsp.ee.ic.ac.uk/sap/

ABSTRACT

We propose a noise estimation algorithm for single channel speech enhancement in highly non-stationary noise environments. The algorithm models time-varying noise using a Hidden Markov Model and tracks changes in noise characteristics by a sequential model update procedure that incorporates a forgetting factor. In addition the algorithm will when necessary create new model states to represent novel noise spectra and will merge existing states that have similar characteristics. We demonstrate that the algorithm is able to track non-stationary noise effectively and show that, when it is incorporated into a standard speech enhancement algorithm, it results in enhanced speech with an improved PESQ score and lower residual noise.

1. INTRODUCTION

Almost all speech enhancement algorithms require an estimate of the noise power spectrum or its equivalent [1, 3]. The accuracy of this estimate has a major impact on the overall quality of the speech enhancement: overestimating the noise will lead to distortion of the speech, while underestimating it will lead to unwanted residual noise.

The noise power spectrum is commonly assumed to change only slowly with time so that it can be estimated as a recursive average. Early systems controlled the averaging process by using a voice-activity detector (VAD) [12] to identify noise-dominated frames. To avoid the VAD requirement, [7, 8] estimate the noise spectrum by taking the minimum of the temporally smoothed power spectrum in each frequency bin and then applying a bias compensation factor. This method is effective in estimating both stationary and time-varying noise even when speech is present but, because it relies on temporal averaging, it is unable to follow abrupt changes in the noise spectrum.

In a realistic environment, especially when using a mobile device, the noise normally includes multiple components. These can vary rapidly due to relative motion between source and receiver or because the sound sources themselves are intermittent (e.g. ringing phones or door slams). Several authors have recognised that such non-stationary noise environments are better modelled as a set of discrete states than as a single time-varying source. In this approach, each state corresponds to a distinct noise power spectrum and the state sequence over time is conveniently represented by a Hidden Markov Model (HMM). In [11] the appropriate noise model, together with its overall gain, is selected during non-speech intervals from a library of pre-trained HMMs. This approach was extended in [15] to include a joint estimate of the time-varying noise and speech gains in every frame. The authors argue that the effect of relative motion between source and receiver is well modelled by a fixed spectral shape with a time-varying gain. A subsequent paper [14] avoids the need for a library of pre-trained noise models by updating the power spectra of the noise states continually using a recursive estimation-maximization (EM) procedure [13, 6]. To protect against divergence of the adaptive algorithm, the state with the lowest occupation probability over recent frames is periodically replaced by a "safety-net state" whose power spectrum is determined using a minimum statistics algorithm [7] provided that this increases the likelihood of the model.

In this paper, we address the problem of estimating highly non-stationary noise environments that include abrupt changes in spectral characteristics. We present an algorithm that uses an HMM to model the noise and that, like [14], tracks slowly evolving noise spectra with a recursive EM update. Unlike previous approaches however, the algorithm is able to detect the presence of a novel noise spectrum and to create a new HMM state to represent it. The algorithm is thus able to follow noise environments with both slowly changing and intermittent components.

In Sec. 2 we present the noise estimation algorithm and the update procedures used for both slowly evolving and abruptly changing noise environments. In Sec. 3 we evaluate the algorithm's performance both in estimating the noise spectrum and when used with a speech enhancement algorithm. Finally in Sec. 4 we present conclusions.

2. PROPOSED NOISE ESTIMATION ALGORITHM

In this paper, we use an HMM to model the noise power spectrum $O_t(k)$ at time frame t and frequency index $k \in \{1 \ldots K\}$. Following [3] we assume that the spectral component of the noise, $o_t(k)$, is Gaussian distributed with uncorrelated real and imaginary parts. Under this assumption, the power spectral components $O_t(k) = |o_t(k)|^2$ will follow a negative exponential distribution,

$$p(O_t(k)) = \frac{1}{E\{O_t(k)\}} \exp\left(-\frac{O_t(k)}{E\{O_t(k)\}}\right) \qquad (1)$$

where $E\{\}$ denotes expectation. With respect to a mean power spectrum μ_0, the log observation probability $\log b(O_t)$ is given by

$$\begin{aligned} \log b(O_t \mid \mu_0) &= \log\left(\prod_k \frac{1}{\mu_0(k)} \exp\left(-\frac{O_t(k)}{\mu_0(k)}\right)\right) \\ &= \sum_k (-\log \mu_0(k) - \frac{O_t(k)}{\mu_0(k)}) \end{aligned} \qquad (2)$$

under the assumption that the frequency components of O_t are conditionally independent given μ_0.

2.1 Hidden Markov Model with recursive updating

The model parameter set for an HMM with M states is $\zeta = (\pi, A, B)$ where $\pi = \{\pi_i\}$ is the vector of initial state probabilities, $A = \{a_{ij}\}$ is the matrix of state transition probabilities and $B = \{b_j(O_t)\}$ is the vector of observation probabilities within each state j. We assume that π is equal to the stationary state probability given by the eigenvector satisfying $A^T\pi = \pi$ and the observations probabilities are determined from (2) using the mean power spectrum, μ_j. Thus we can redefine the noise model as $\zeta = \{\mu, A\}$ where $\mu = \{\mu_j\}$. We will use a $^{(T)}$ superscript to denote the model parameters estimated from the observations $O^{(T)} = \{O_t : t \in [1,T]\}$ (e.g. $\zeta^{(T)}$, $\mu_i^{(T)}$, ...) but we will normally omit the superscript if all quantities in an equation relate to the same model.

2.1.1 Model update with forgetting factor

We assume in this section that the noise characteristics are slowly evolving; tracking rapidly changing noise spectra will be addressed in Sec. 2.2. We can derive the forward and backward state probabilities, $\alpha_i(t)$ and $\beta_i(t)$ from the model, $\zeta^{(T)}$, and the observations, $O^{(T)}$,

$$\alpha_i(t) = \sum_j \alpha_j(t-1)a_{ji}b_i(O_t) \quad (3)$$
$$\text{with} \quad \alpha_i(0) = \pi_i$$
$$\beta_i(t) = \sum_j a_{ij}b_j(O_{t+1})\beta_j(t+1) \quad (4)$$
$$\text{with} \quad \beta_i^{(T)}(T) = \pi_i$$

where $b_j(O_t)$ is the observation probability of O_t belonging to the state j as given by (2). In estimating the model, we would like to weight the recent frames more strongly than frames in the distant past. Accordingly we introduce a forgetting factor, λ; a similar approach was taken in [6]. The model update equations are

$$\mu_i^{(T)} = \frac{U_i^{(T)}(1,T)}{Q_i^{(T)}(1,T)}$$
$$a_{ij}^{(T)} = \frac{a_{ij}^{(T-1)}R_{ij}^{(T)}(1,T-1)}{Q_i^{(T)}(1,T-1)} \quad (5)$$

where U, Q and R are defined in (6)-(8) below, with the superscript T omitted for clarity

$$U_i(\tau_1,\tau_2) = \frac{1}{P}\sum_{t=\tau_1}^{\tau_2} \lambda^{\tau_2-t}\alpha_i(t)\beta_i(t)O_t \quad (6)$$
$$Q_i(\tau_1,\tau_2) = \frac{1}{P}\sum_{t=\tau_1}^{\tau_2} \lambda^{\tau_2-t}\alpha_i(t)\beta_i(t) \quad (7)$$
$$R_{ij}(\tau_1,\tau_2) = \frac{1}{P}\sum_{t=\tau_1}^{\tau_2} \lambda^{\tau_2-t}\alpha_i(t)b_j(O_{t+1})\beta_j(t+1) \quad (8)$$

where $P^{(T)} = \sum_i \alpha_i^{(T)}(T)\beta_i^{(T)}(T)$ is the probability of generating $O^{(T)}$. These are the standard Baum-Welch update equations [9] except for the exponential factor λ^{T-t}.

2.1.2 Time-update

Assuming now that we have determined $\zeta^{(T-1)}$ and now wish to update the model to time ζ^T. Re-evaluating (3)-(5) directly would require us to save the entire set of observations $\{O_t\}$. To avoid this, we retain only the L most recent observations and assume that for sufficiently old frames, the state occupation probabilities are unchanged, i.e.

$$\frac{\alpha_i^{(T)}(t)\beta_i^{(T)}(t)}{P^{(T)}} \approx \frac{\alpha_i^{(T-1)}(t)\beta_i^{(T-1)}(t)}{P^{(T-1)}} \quad \text{for} \quad t \le T-L.$$

With this assumption, and writing $T_L = T - L$ for compactness, we can write the update equations as

$$\mu_i^{(T)} \approx \frac{U_i^{(T-1)}(1,T_L) + U_i^{(T)}(T_L+1,T)}{Q_i^{(T-1)}(1,T_L) + Q_i^{(T)}(T_L+1,T)} \quad (9)$$

$$a_{ij}^{(T)} \approx \frac{a_{ij}^{(T-1)}\left(R_{ij}^{(T-1)}(1,T_L-1) + R_{ij}^{(T)}(T_L,T-1)\right)}{Q_i^{(T-1)}(1,T_L-1) + Q_i^{(T)}(T_L,T-1)}. \quad (10)$$

The advantage of these expressions is that the first terms in the numerator and denominator of both (9) and (10) can be calculated recursively without reference to past observations and the sums implicit in the second terms extend over only the past L observations. To update the model, we initialize

$$\mu_i^{(T)} = \mu_i^{(T-1)}$$
$$a_{ij}^{(T)} = a_{ij}^{(T-1)}$$
$$\alpha_i^{(T)}(T-1) = \alpha_i^{(T-1)}(T-1),$$

and calculate $\alpha_i(T)$ from (3), $\beta_j(t)$ for $t \in [T_L+1, T]$ from (4) and $P^{(T)}$. We can then calculate all the remaining quantities in (9) and (10) and update the model. Finally, in preparation for the next time step, we update the first terms in the numerator and denominator of (9) and (10) using

$$U_i^{(T)}(1,T_L+1) = \lambda U_i^{(T-1)}(1,T_L) + \frac{\alpha_i^{(T)}(T_L+1)\beta_i^{(T)}(T_L+1)O_{T-L+1}}{P^{(T)}}$$
$$Q_i^{(T)}(1,T_L+1) = \lambda Q_i^{(T-1)}(1,T_L) + \frac{\alpha_i^{(T)}(T_L+1)\beta_i^{(T)}(T_L+1)}{P^{(T)}}$$
$$R_{ij}^{(T)}(1,T_L) = \lambda R_{ij}^{(T-1)}(1,T_L-1) + \frac{\alpha_i^{(T)}(T_L-1)b_j^{(T)}(O_{T_L})\beta_i^{(T)}(T_L)}{P^{(T)}} \quad (11)$$

2.1.3 Model initialization

We initialize the model conventionally from the first T_0 frames where $T_0 \gg M$. We first cluster the T_0 observation vectors into M states and then use Viterbi training [9], modified to include the forgetting factor λ, to create an initial model, $\zeta^{(T_0)} = \left\{\mu^{(T_0)}, A^{(T_0)}\right\}$.

2.2 Adapting to fast changing noise characteristics

In order to accommodate an abrupt change to the noise characteristics as might, for example, arise from the introduction of a novel noise source, we need to create a new state to model the newly observed noise spectrum. At the same time, to avoid increasing the total number of states, we need to merge two of the existing states. In order to decide when to introduce a new state, we calculate a measure $Z^{(T)}$ that indicates how well the most recent L frames of observed data fit the current model, $\zeta^{(T)}$. From (2), it is straightforward to show that if $E\{O_t\} = \mu$, then

$$\begin{aligned} E\{\log b(O_t \mid \mu)\} &= -\sum_k (\log \mu(k) + 1) \\ Var\{\log b(O_t \mid \mu)\} &= K \end{aligned}$$

Accordingly we define $Z^{(T)}$ as the normalized difference between the weighted log-likelihood of the most recent L frames and its expectation

$$Z^{(T)} = \frac{\sum_{t=T_L+1}^{T} \lambda^{T-t} \sum_k \left(1 - \frac{O_t(k)}{\mu_{i(t)}(k)}\right)}{\sqrt{K \sum_{t=T_L+1}^{T} (\lambda^{T-t})^2}} \tag{12}$$

where $i(t)$ gives the state occupied at time t in the maximum likelihood state sequence.

If $\left|Z^{(T)}\right|$ exceeds an empirically determined threshold, θ_Z, then this indicates that $\zeta^{(T)}$ should be re-estimated and a new type of noise might be present. In this case, we therefore create a tentative model, $\hat{\zeta}^{(T)}$, in which two of the existing states are merged and a new state created. For the tentative model $\hat{\zeta}^{(T)}$, we first determine the pair of states, $\{i, j\}$, whose merging will cause the least reduction in likelihood. We then initialize the state means for the model as

$$\hat{\mu}_r^{(T-1)} = \begin{cases} O_T & \text{for } r = j \\ \frac{Q_i(1,T_L)\mu_i^{(T-1)} + Q_j(1,T_L)\mu_j^{(T-1)}}{Q_i(1,T_L) + Q_j(1,T_L)} & \text{for } r = i \\ \mu_r^{(T-1)} & \text{otherwise} \end{cases}$$

Thus state j models the new noise spectrum (which we assume is exemplified in frame T) and state i is initialized as a weighted average of the previous states i and j. We re-train this initial model, $\hat{\zeta}^{(T-1)}$, using Viterbi training on the most recent L frames, $\{O_t : t \in [T_L+1, T]\}$, and re-calculate the accumulated sums by distributing them to each of the new states according to the new mean $\hat{\mu}^{(T-1)}$:

$$\begin{aligned} \hat{U}_i^{(T-1)}(1,T_L) &= \sum_m \phi_{mj} U_m^{(T-1)}(1,T_L) \\ \hat{Q}_j^{(T-1)}(1,T_L) &= \sum_m \phi_{mj} Q_m^{(T-1)}(1,T_L) \\ \hat{R}_{ij}^{(T-1)}(1,T_L-1) &= \sum_m \sum_n \phi_{mi}\phi_{nj} R_{mn}^{(T-1)}(1,T_L-1) \end{aligned} \tag{13}$$

where ϕ_{ij} estimate the probability of a frame that was previously in state i being in state j of the new model, $\phi_{ij} = \frac{b\left(\mu_i^{(T-1)} \mid \hat{\mu}_j^{(T-1)}\right)}{\sum_j b\left(\mu_i^{(T-1)} \mid \hat{\mu}_j^{(T-1)}\right)}$. Using the EM re-estimation algorithm from (9) & (11), $\hat{\zeta}^{(T)}$ is obtained. However, we only wish to use this revised model if it will result in an increase in log likelihood. Accordingly the increase, $I^{(T)}$, in the log-likelihood is estimated as

$$\begin{aligned} I^{(T)} = & \sum_{t=T_L+1}^{T} \lambda^{T-t} \sum_i Q_i(t,t) \log b(O_t, \hat{\mu}_i) - \\ & \sum_{t=T_L+1}^{T} \lambda^{T-t} \sum_i Q_i(t,t) \log b(O_t, \mu_i) - \\ & \frac{\lambda^L}{1-\lambda} \sum_i \sum_j \phi_{ij} \pi_i D(\mu_i, \hat{\mu}_j) \end{aligned} \tag{14}$$

where $D(\mu_i, \hat{\mu}_j) = \sum_k \left(\frac{\mu_i(k)}{\hat{\mu}_j(k)} - \log \frac{\mu_i(k)}{\hat{\mu}_j(k)} - 1\right)$ is the Itakura-Saito distance and equals the expected increase in log likelihood of a frame whose true mean power spectrum is μ_i is modeled by a state with mean $\hat{\mu}_j$. The first two terms in (14) give the log likelihood improvement over the most recent L frames while the last term approximates the decrease in log likelihood of the earlier frames.

2.3 Noise estimation algorithm overview

In this section, we outline the processing steps of the proposed algorithm as follows:

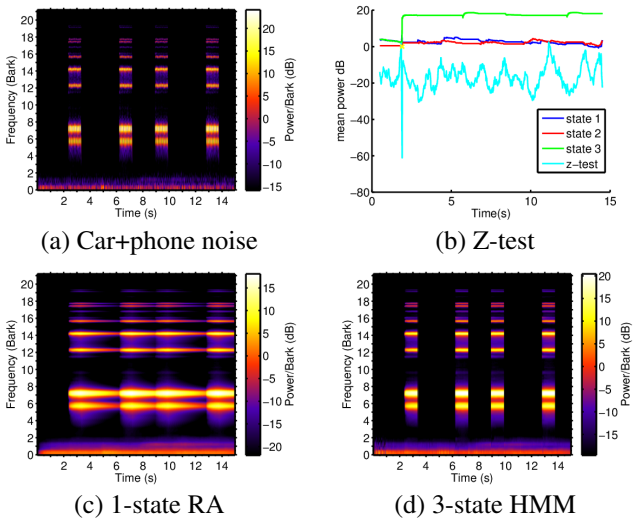
1. Compute the initialized model $\zeta^{(T_0)}$ using Viterbi training and set $T = T_0$.
2. Compute and update the model $\zeta^{(T)}$ using (9) - (11).
3. Compute the $Z^{(T)}$ using (12).
4. If $Z^{(T)} > \theta_Z$, re-train the model $\hat{\zeta}^{(T-1)}$ using parameters described in (13), else skip to step 7.
5. Compute $I^{(T)}$ using (14).
6. If $I^{(T)} > 0$, update the model $\zeta^{(T)} = \hat{\zeta}^{(T)}$.
7. Increment $T = T + 1$, and go back to step 2 for the next time frame.

2.4 Noise Estimator during Speech activity

We assume an external voice activity detector (VAD) and only update the noise model when speech is absent. During speech presence we freeze the noise model ζ, and use it to estimate the noise state for each frame as follows. We assume that the clean speech power spectrum may be approximated as $\gamma_t \bar{\mu}$ where $\bar{\mu}$ is the Long-Term Average Speech Spectrum (LTASS) [4] and γ_t is the speech level at time t. For each noise state, j, we evaluate the likelihood $b(O_t \mid \mu_j + \gamma_t \bar{\mu})$ and select the γ_t that maximizes it; the observation probabilities are therefore given by $\max_{\gamma_t} b(O_t \mid \mu_j + \gamma_t \bar{\mu})$. Once we have evaluated the observation probabilities we can use the Viterbi algorithm to determine the most likely noise state sequence. The noisy speech is then enhanced using the MMSE algorithm [3] using the corresponding noise state means, μ_j, as the a priori noise estimates. It is possible to impose temporal continuity constraints on γ_t, but we have not found that this gives a significant improvement in noise state estimation.

3. PERFORMANCE EVALUATION

In this section, we first demonstrate the noise tracking abilities of our proposed multi-state HMM noise estimation algorithm. Then in the context of the speech enhancement,



(a) Car+phone noise (b) Z-test

(c) 1-state RA (d) 3-state HMM

Figure 1: Spectrogram of (a) car+phone noise, with its estimation using (c) 1-state recursive averaging (d) a 3-state HMM; (b) Power of estimated noise states and the Z-test value

we compare the performance of our noise algorithm with other noise estimation algorithms. All signals are sampled at a frequency of 16 kHz. The time-frames have a length of 32 ms with a 50% overlap resulting in $K = 257$ frequency bins. We retain the most recent $L = 30$ frames (480 ms), and also set $T_0 = 30$. The forgetting factor is chosen to be $\lambda = 1 - 1/(2L)$, which gives a time constant of $2L = 960$ ms The same value of λ is used for the 1-state recursive averaging model in which $\mu^{(T)} = (1-\lambda)\,\mu^{(T-1)} + \lambda O_T$. The threshold θ_Z defined in Sec. 2.2 is set to 30.

3.1 Noise Estimation

In this experiment, the noise of a ringing phone is added to a background car engine noise which is predominantly low frequency. Fig. 1(a) shows the spectrogram of this composite noise and it can be seen that the noise spectrum changes abruptly whenever the phone rings. The spectrogram of the estimated noise using 1-state recursive averaging (RA) method is shown in Fig. 1(c); this is representative of noise estimators that assumes a quasi-stationary noise spectrum. As would be expected this model is unable to track the rapidly changing noise and smears the spectrum in the time direction. A 3-state HMM model is used to estimate this noise, and the three higher time waveforms in Fig. 1(b) show the mean power of each state. The scaled value of $Z^{(T)}$ with an offset of -20, which measures how well the L most recent observations fit the model, is plotted below the waveforms. We see that when the first phone ring occurs, at approximately 2.3 s, there is an abrupt fall in $Z^{(T)}$ which indicates the arrival of a novel noise spectrum. Two of the existing states, state 2 and 3, are therefore merged and state 3 is reallocated to model the new noise spectrum. The corresponding spectrogram for our proposed model is shown in Fig. 1(d) in which the estimated noise spectrum follows the state mean of the maximum likelihood state sequence. We see that the abrupt changes in noise spectrum are perfectly tracked and both noises are well modelled.

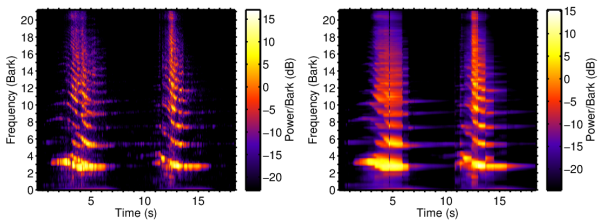
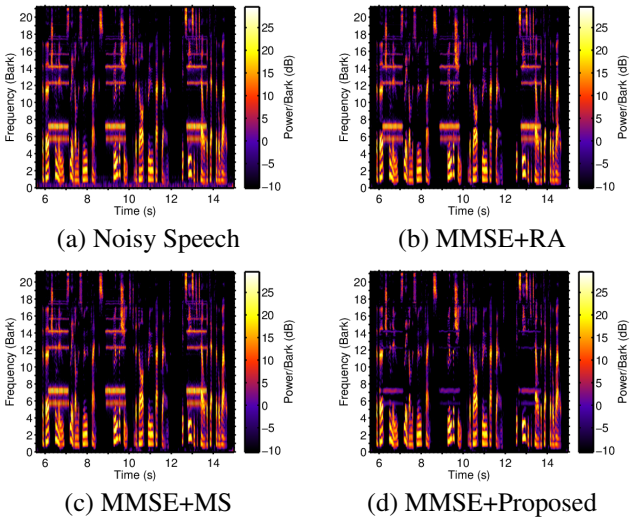


Figure 2: Spectrogram of (a) Formula 1 noise and (b) its estimate using a 8-state HMM



(a) Noisy Speech (b) MMSE+RA

(c) MMSE+MS (d) MMSE+Proposed

Figure 3: Spectrogram of (a) the unenhanced noisy speech corrupted by the car+phone noise at 20 dB SNR, and the MMSE enhanced speech using different noise estimator (b) RA (c) MS (d) HMM

Fig. 2 (a) shows the corresponding spectrogram for the more complex sound of a Formula 1 race. As the vehicles pass the microphone their Doppler-shifted engine pitch falls and we see from the spectrogram's in Fig. 2(b) is able to preserve some details of this non-stationary spectrum.

3.2 Speech Enhancement

Fig. 3(a) shows an example of a speech signal corrupted by the car+phone noise from Fig. 1(a) at 20 dB SNR. The first 5 seconds of the signal contain no speech and are used to train the model; this noise-only segment is not included in the spectrogram.

The noisy speech signal is enhanced by the MMSE algorithm [3] using three different noise estimators. Fig. 3 shows the enhanced speech signals using respectively (b) 1-state recursive averaging (RA), (c) minimum statistics (MS) [2] and (d) multi-state hidden Markov model (HMM). The RA and HMM estimators are trained on the initial noise-only segment and frozed roughly at $t = 5$ seconds, while the MS estimator is allowed to adapt continuously throughout the signal. We see that the stationary low frequency noise component is effectively removed using all three methods but only with the HMM method is the phone ringing largely eliminated. As seen in Fig. 1(c), the noise estimate from the RA method is blurred in time and so, with this estimate, distor-

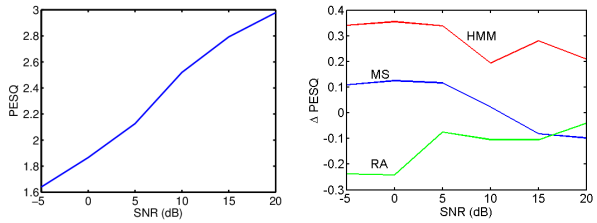


Figure 4: (a) average PESQ scores of unenhanced noisy speech (b) average improvement of PESQ scores at different SNRs

tion is introduced in the gaps between rings. Even though the MS method tracks the variation of noise level during speech presence, it cannot cannot respond quickly enough to eliminate the phone noise. Even though the training segment includes only one phone ring, this is sufficient for the HMM method to learn its characteristics and to attenuate it greatly when it subsequently occurs. We assess the quality of the speech by means of the PESQ (Perceptual Evaluation of Speech Quality) score [5]. The PESQ score for the unenhanced noisy speech in Fig. 3(a) is 2.55. The improvement of the PESQ score between the noisy and enhanced speech when using the RA, MS and HMM methods to estimate the noise is respectively -0.17, -0.07 and 0.43, indicating that our proposed HMM method gives a noticeably greater quality improvement than the other methods.

A second set of experiments was performed with noise+speech at different SNRs using, as before, a noise-only segment at the beginning of the signal for training the model. 10 different clean speech signals were chosen from IEEE sentence database [10], of average duration about 10s. The noises used were the car+phone and Formula 1 noise used above. Three different noise estimation algorithms were evaluated: (i) 1-state recursive averaging (RA), (ii) minimum statistics (MS) and (iii) our proposed 8-state addaptive hidden Markov model (HMM). Fig. 4(a) shows the average PESQ score as a function of SNR and Fig. 4(b) shows the improvement obtained when the MMSE enhancer is used with each of the noise estimators. We see that the RA estimate degrades the PESQ score at all SNRs while the MS estimate gives a small improvement at low SNRs. In contrast, the proposed improves the PESQ score at all SNRs and consistently outperforms the other methods.

4. CONCLUSION

We have proposed a continuous noise estimator using the HMM with recursive EM algorithm. The proposed algorithm could adaptively update the model parameters with the change of the noise level and spectral characteristics. We have also proposed a noise power estimator from the noisy speech given the noise model. We showed through objective evaluations that our proposed method achieves a more accurate noise model for highly nonstationary noise with abrupt changing of noise sources. Thus when incorporating into a speech enhancement system, it gives a better speech quality by facilitating a lower level of residual noise. We are currently extending our work to update the noise model during speech activity.

REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[2] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," 1997. [Online]. Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[4] *Artificial Voices*, International Telecommunications Union (ITU-T) Recommendation P.50, Sep. 1999.

[5] *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, International Telecommunications Union (ITU-T) Recommendation P.862, Feb. 2001.

[6] V. Krishnamurthy and J. B. Moore, "On-line estimation of hidden Markov model parameters based on the Kullback-Leibler information measure," *IEEE Trans. Signal Process.*, vol. 41, no. 8, pp. 2557–2573, 1993.

[7] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 504–512, Jul. 2001.

[8] ——, "Bias compensation methods for minimum statistics noise power spectral density estimation," *Signal Processing*, vol. 86, no. 6, pp. 1215–1229, Jun. 2006.

[9] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[10] E. H. Rothauser, W. D. Chapman, N. Guttman, M. H. L. Hecker, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, no. 3, pp. 225–246, 1969.

[11] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 445–455, Sep. 1998.

[12] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[13] D. Titterington, "Recursive parameter estimation using incomplete data," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 46, no. 2, pp. 257–267, 1984.

[14] D. Y. Zhao, W. B. Kleijn, A. Ypma, and B. de Vries, "Online noise estimation using stochastic-gain HMM for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 835–846, 2008.

[15] D. Y. Zhao and W. B. Kleijn, "Hmm-based gain modeling for enhancement of speech in noise," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 882–892, 2007.