SPARSE HMM-BASED SPEECH ENHANCEMENT METHOD FOR STATIONARY AND NON-STATIONARY NOISE ENVIRONMENTS

Feng Deng[1], *Chang-chun Bao*[1], *and W. Bastiaan Kleijn*[1 2]

[1]Speech and Audio Signal Processing Laboratory, School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing, China
[2]School of Engineering and Computer Science, Victoria University of Wellington, New Zealand
E-mail: dengfeng@emails.bjut.edu.cn, baochch@bjut.edu.cn, bastiaan.kleijn@ecs.vuw.ac.nz

ABSTRACT

We propose a sparse hidden Markov model (HMM)-based single-channel speech enhancement method that models the speech and noise gains accurately in both stationary and non-stationary environments. The objective function is augmented with an l_p regularization term resulting in a sparse autoregressive HMM (SARHMM). The method encourages sparsity in the speech- and noise- modeling, which eliminates the ambiguity between noise and speech spectra and, as a consequence, provides improved tracking of the changes of both spectral shapes and power levels of non-stationary noise. Using the modeled speech and noise SARHMMs, we first construct an estimator to estimate the noise spectrum. Then a Bayesian speech estimator is used to obtain the enhanced speech. The test results indicate that the proposed speech enhancement scheme performs much better than the reference methods in non-stationary environments, while providing state-of-the-art performance for stationary conditions.

Index Terms— Speech Enhancement, Sparse ARHMM, Gain Modeling, Non-stationary Noise

1. INTRODUCTION

Traditional speech enhancements, such as Wiener filtering [1], the spectral-subtraction method [2] and the MMSE method of Ephraim-Malah[3], generally do not perform well in non-stationary noise environments as it is not possible to specify prior knowledge about the speech and noise. To address speech enhancement in non-stationary environments, auto-regressive HMM(ARHMM)[4] and trained codebooks [5][6][7] have been used successfully to model the statistics of speech and noise. These methods model the change of speech and noise spectral characteristics (i.e. spectral envelope or spectral shapes).

As the accurate modeling of the gain variances of the speech and noise can play an important role in speech enhancement in non-stationary noise environments, a revised ARHMM-based method was proposed in [8]. In this approach, the speech and noise gains are considered as random processes that describe the power levels of speech and noise, respectively. The characteristics of speech and noise are learned online, facilitating an accurate prior knowledge of the gains. Motivated by its good performance, we used [8] as the basis for the work presented in this present paper.

A problem with the Bayesian estimation procedures used to estimate the speech in the ARHMM-based approach is that the linear combinations of multiple ARHMM states that are used to describe observed speech spectral shapes are not restricted to look like speech. For example, two spectra with three formants may combine to a spectrum with six formants that cannot be produced by a human. Such combined speech spectra can be used to represent a noise observation. Similarly, noise spectra can be combined to represent speech spectra. Therefore, an *ambiguity problem* between the spectral shapes of speech and those of noise exists, which limits the overall performance of the ARHMM gain model.

In this paper, we propose a solution to the ambiguity problem by introducing sparsity into the ARHMM model and derive a new approach that we refer to as *sparse* ARHMM (SARHMM). In the SARHMM, the sparsity is induced to the transition probabilities and the observation probabilities by using an l_p regularization [9], which ensures that only a few states have a significant contribution to the modeled signal for any segment. The sparsity of speech and noise modeling helps to improve the tracking of the changes of both spectral shapes and power levels of non-stationary noise.

ARHMM-based methods [4] [8], have a second inherent problem: clearly audible noise remains between the harmonics for the estimated voiced speech. We present a new solution that addresses this problem. We exploit the fact that the SARHMM provides estimates of both the noise and speech parameters, which aid in finding of a good noise model estimate. Therefore, instead of estimating speech directly, we first construct a noise estimator to estimate the noise power spectrum. Then, a Bayesian speech estimator [10] is derived to obtain the enhanced speech, in which the noise between the harmonics of voiced speech has been removed.

2. THE ARHMM SIGNAL MODEL

Assuming that the clean speech $\boldsymbol{X}_t$ is contaminated by an uncorrelated additive noise $\boldsymbol{W}_t$, then the t'th frame noisy speech $\boldsymbol{Y}_t$ (K samples) can be modeled as: $\boldsymbol{Y}_t = \boldsymbol{X}_t + \boldsymbol{W}_t$.

Let $\boldsymbol{x}_0^{T-1} = \{\boldsymbol{x}_0,...,\boldsymbol{x}_{T-1}\}$ denote the realization of a clean speech sequence from frame 0 to T-1. In the following, we use the overbar '$\bar{}$' to label the parameters as belonging to the speech ARHMM. The probability density function (pdf) of $\boldsymbol{x}_0^{T-1}$ are modeled by an $\bar{N}$-state ARHMM as [4][8].

$$p(\boldsymbol{x}_0^{T-1}) = \sum_{\bar{s}_0^{T-1}} \prod_{t=0}^{T-1} \bar{a}_{\bar{s}_{t-1}\bar{s}_t} p_{\bar{s}_t}(\boldsymbol{x}_t), \tag{1}$$

where $\bar{s}_0^{T-1} = (\bar{s}_t)_{t=0}^{T-1}$ denotes a sequence of speech ARHMM states, and $\bar{s}_t \in \{1,...,\bar{N}\}$ denotes the state of speech at frame t, $\bar{a}_{\bar{s}_{t-1}\bar{s}_t}$ is the state transition probability from state $\bar{s}_{t-1}$ to $\bar{s}_t$, $\bar{a}_{\bar{s}_{-1}\bar{s}_0}$ is the probability of the initial state $\bar{s}_0$. The dependency of the probability density of the clean speech on the speech gain can be made explicit by means of the law of total probability [8]:

$$p_{\bar{s}_t}(\boldsymbol{x}_t) = \int_{-\infty}^{\infty} p_{\bar{s}_t}(\bar{g}_t^{'}) p_{\bar{s}_t}(\boldsymbol{x}_t|\bar{g}_t^{'}) d\bar{g}_t^{'}, \tag{2}$$

where $\bar{g}_t^{'} = \log(\bar{g}_t)$, and $\bar{g}_t$ denotes the linear speech gain. Following [8], we model the pdf $p_{\bar{s}_t}(\bar{g}_t^{'})$ of $\bar{g}_t$ as a state-dependent log-normal distribution:

$$p_{\bar{s}_t}(\bar{g}_t^{'}) = \frac{1}{\sqrt{2\pi\bar{\sigma}_{\bar{s}_t}^2}} \exp\left(- \frac{\left[\bar{g}_t^{'} - (\bar{\mu}_{\bar{s}_t} + \bar{q}_t)\right]^2}{2\bar{\sigma}_{\bar{s}_t}^2} \right), \tag{3}$$

where $\bar{\mu}_{\bar{s}_t} + \bar{q}_t$ denotes a mean value and $\bar{\sigma}_{\bar{s}_t}^2$ denotes the variance. The parameters $\bar{\mu}_{\bar{s}_t}$ and $\bar{\sigma}_{\bar{s}_t}^2$ are time-invariant and can be estimated off-line. The parameter $\bar{q}_t$ is used to compensate for the speech-gain bias, which can be estimated on-line.

For a given gain $\bar{g}_t$, we assume speech to be a zero-mean $\bar{p}$'th order Gaussian AR processes. The pdf $p_{\bar{s}_t}(\boldsymbol{x}_t|\bar{g}_t^{'})$ given a speech gain $\bar{g}_t^{'}$ can be described as [4][8]

$$p_{\bar{s}_t}(\boldsymbol{x}_t|\bar{g}_t^{'}) = \frac{\exp\left(- \frac{1}{2\bar{g}_t} \boldsymbol{x}_t^{\#} \bar{\boldsymbol{D}}_{\bar{s}_t}^{-1} \boldsymbol{x}_t \right)}{(2\pi\bar{g}_t)^{K/2} |\bar{\boldsymbol{D}}_{\bar{s}_t}|^{1/2}}, \tag{4}$$

where $^{\#}$ denotes Hermitian transposition, $|.|$ denotes the determinant, $\bar{\boldsymbol{D}}_{\bar{s}_t} = (\bar{\boldsymbol{A}}_{\bar{s}_t}^{\#} \bar{\boldsymbol{A}}_{\bar{s}_t})^{-1}$ is the covariance matrix of the AR process, $\bar{\boldsymbol{A}}_{\bar{s}_t}$ is a $K \times K$ lower triangular Toeplitz matrix in which the first $\bar{p}+1$ elements of the first column constitute the speech AR coefficients $(\bar{\alpha}_0, \bar{\alpha}_1, ..., \bar{\alpha}_{\bar{p}})$.

To capture the high diversity and variability of acoustical noises in a non-stationary environment, we model the noise similarly to the speech. Thus, we use a noise ARHMM that is nearly identical in form to the ARHMM for the speech. We label the noise model parameters by '$\ddot{}$', in contrast to the overbar '$\bar{}$' for the speech model. The pdf $p_{\ddot{s}_t}(\boldsymbol{w}_t)$ is defined similarly to equations (1). However, for noise we use only a single gain [8]. Thus, the noise gain is modeled as

$$p(\ddot{g}_t^{'}) = \frac{1}{\sqrt{2\pi\ddot{\sigma}^2}} \exp\Big(- \frac{(\ddot{g}_t^{'} - \ddot{\mu}_t)^2}{2\ddot{\sigma}^2} \Big), \tag{5}$$

where the parameters have the same meaning as in the case of (3). The gain-conditional probability density of the noise, $p_{\ddot{s}_t}(\boldsymbol{w}_t|\ddot{g}_t^{'})$ is defined similarly to that for speech in (4).

Based on the speech and noise ARHMM models, we can derive the pdf of the noisy speech. Let $s_t = (\bar{s}_t, \ddot{s}_t)$ denote the noisy speech state at frame t, which is a composite state of speech and noise, so there are $\bar{N} \times \ddot{N}$ states in the noisy speech model. The pdf $p_{s_t}(\boldsymbol{y}_t)$ of noisy speech $\boldsymbol{y}_t$ for a given composite state s_t can be written as [8]:

$$\begin{aligned} p_{s_t}(\boldsymbol{y}_t) &= \int\int p_{s_t}(\boldsymbol{y}_t, \bar{g}_t^{'}, \ddot{g}_t^{'}) d\bar{g}_t^{'} \ddot{g}_t^{'} \\ &= \int\int p_{\bar{s}_t}(\bar{g}_t^{'}) p(\ddot{g}_t^{'}) p_{s_t}(\boldsymbol{y}_t|\bar{g}_t^{'}, \ddot{g}_t^{'}) d\bar{g}_t^{'} \ddot{g}_t^{'}, \end{aligned} \tag{6}$$

where the pdf $p_{s_t}(\boldsymbol{y}_t|\bar{g}_t^{'}, \ddot{g}_t^{'})$ is a Gaussian distribution with zero-mean and covariance $\boldsymbol{D}_{s_t} = \bar{g}_t \bar{\boldsymbol{D}}_{\bar{s}_t} + \ddot{g}_t \ddot{\boldsymbol{D}}_{\ddot{s}_t}$.

By using a scaled Dirac delta function $\delta(.)$, the integral in (6) can be approximated as [8]:

$$\begin{aligned} p_{s_t}(\boldsymbol{y}_t) &\approx p_{s_t}(\boldsymbol{y}_t, \hat{\bar{g}}_t^{'}, \hat{\ddot{g}}_t^{'}) \\ \{\hat{\bar{g}}_t^{'}, \hat{\ddot{g}}_t^{'}\} &= \arg\max_{\bar{g}_t^{'}, \ddot{g}_t^{'}} \log p_{s_t}(\boldsymbol{y}_t, \bar{g}_t^{'}, \ddot{g}_t^{'}), \end{aligned} \tag{7}$$

The validity of this approximation was confirmed in [8], and the concrete procedure to obtain the optimal speech and noise gain pair $\{\hat{\bar{g}}_t^{'}, \hat{\ddot{g}}_t^{'}\}$ was also discussed there.

3. PARAMETER ESTIMATION OF SARHMMS

To solve the ambiguity problems, we proposed a sparse ARHMM (SARHMM), and in this section, the off-line estimation of the time invariant parameters of the speech and noise SARHMM are considered. The on-line estimate of the time-varying parameters is similar to that in [8].

The ARHMM parameters are often trained by the Baum-Welch approach that is based on the expectation maximization (EM) algorithm [11]. The auxiliary Q function of speech model often can be split into separate terms for the three types of model parameters $\theta = (\pi, \boldsymbol{A}, \boldsymbol{B})$: the initial distribution of states (π), the state transition probability matrix ($\boldsymbol{A}$) and the observation probability matrix ($\boldsymbol{B}$).

$$\begin{aligned} Q(\theta, \theta^{'}) &= \sum_{\bar{s}_t} \log p(\bar{s}_t|O,\theta) p(\bar{s}_t|O,\theta^{'}) \\ &= \sum_{\bar{s}_t} \log \pi_{\bar{s}_0} p(\bar{s}_t|O,\theta^{'}) + \sum_{\bar{s}_t} \sum_{t=0}^{T-1} \log \bar{a}_{\bar{s}_{t-1}\bar{s}_t} p(\bar{s}_t|O,\theta^{'}) \\ &\quad + \sum_{\bar{s}_t} \sum_{t=0}^{T-1} \log p_{\bar{s}_t}(\boldsymbol{x}_t) p(\bar{s}_t|O,\theta^{'}), \end{aligned} \tag{8}$$

where $O = (\boldsymbol{x}_t)_{t=0}^{T-1}$ is the observation sequence, $\theta^{'}$ is the previous estimation of model parameters θ.

Following [9], we first encourage sparsity to transition probabilities by introducing the l_p norm $H(\boldsymbol{A}) = ||\boldsymbol{A}||_{1,p_1}$ to the second term of equation (8), and then we can derive the update equation of transition probabilities of SARHMM as

$$\bar{a}_{ij} = \frac{\Big(\sum_{t=0}^{T-1} p(\bar{s}_{t-1} = i, \bar{s}_t = j|O, \theta^{'}) - \eta_1 \bar{A}_{ij}\Big)^{+}}{\sum_{h}^{\bar{N}} \Big(\sum_{t=0}^{T-1} p(\bar{s}_{t-1} = i, \bar{s}_t = h|O, \theta^{'}) - \eta_1 \bar{A}_{ih}\Big)^{+}} \quad (9)$$

where $(\cdot)^{+} = \max(\cdot, 0)$. $\bar{A}_{ij}$ is the regularization term for transition probability, which is defined as

$$\bar{A}_{ij} = \bar{a}_{ij} \nabla_{\bar{a}_{ij}} H(\boldsymbol{A}) = \bar{a}_{ij} \big[\bar{a}_{ij} / (\sum_{h} \bar{a}_{ih}^{p_1})^{1/p_1}\big]^{p_1 - 1}, \quad (10)$$

where ∇ is the gradient operator and p_1 is a regularization parameter.

We also can encourage sparsity to the observation probability $p_{\bar{s}_t}(\boldsymbol{x}_t)$ of speech ARHMM by introducing the l_p norm $H(\boldsymbol{B}) = ||\boldsymbol{B}||_{1,p_2}$ (p_2 is a regularization parameter) to the third term of equation (8). Similar to the derivation of equation (10), the regularization term $\bar{B}_{\bar{s}_t.x_t}$ for observation probability of speech SARHMM can be obtained by

$$\begin{aligned} \bar{B}_{\bar{s}_t.x_t} &= p_{\bar{s}_t}(x_t) \nabla_{p_{\bar{s}_t}(x_t)} H(\boldsymbol{B}) \\ &= p_{\bar{s}_t}(x_t) \left[\frac{p_{\bar{s}_t}(x_t)}{\big(\sum_t p_{\bar{s}_t}^{p_2}(x_t)\big)^{1/p_2}} \right]^{p_2 - 1}, \end{aligned} \quad (11)$$

Using the regularization term $\bar{B}_{\bar{s}_t.x_t}$ to the third term of equation (8), we can derive the update equations of the training parameters $\bar{\theta} = \{\bar{\mu}_{\bar{s}}, \bar{\sigma}_{\bar{s}}^2, \bar{\alpha}_{\bar{s}}, \bar{q}_{\bar{r}}\}$ of observation probability for the *j*th iteration as (12)-(14):

$$\bar{\mu}_{\bar{s}}^{(j)} = \frac{\sum_r \sum_t \hat{\bar{\omega}}(\bar{s}_t) \int \bar{g}_t^{'} \Xi(\bar{s}_t) d\bar{g}_t^{'} - \bar{q}_r}{\sum_r \sum_t \hat{\bar{\omega}}(\bar{s}_t)}, \quad (12)$$

$$\bar{\sigma}_{\bar{s}}^{2(j)} = \frac{\sum_r \sum_t \hat{\bar{\omega}}(\bar{s}_t) \int (\bar{g}_t^{'} - \bar{\mu}_{\bar{s}}^{(j)} - \bar{q}_r)^2 \Xi(\bar{s}_t) d\bar{g}_t^{'}}{\sum_r \sum_t \hat{\bar{\omega}}(\bar{s}_t)}, \quad (13)$$

$$\bar{q}_r^{(j)} = \frac{\sum_r \sum_t \frac{\hat{\bar{\omega}}(\bar{s}_t)}{\bar{\sigma}_{\bar{s}}^{2(j)}} \int (\bar{g}_t^{'} - \bar{\mu}_{\bar{s}}^{(j)}) \Xi(\bar{s}_t) d\bar{g}_t^{'}}{\sum_r \sum_t \frac{\hat{\bar{\omega}}(\bar{s}_t)}{\bar{\sigma}_{\bar{s}}^{2(j)}}}, \quad (14)$$

where $\bar{\omega}(\bar{s}_t) = p(\bar{s}_t|x_0^{T-1}, \hat{\bar{\theta}}^{(j-1)})$ is the posterior state probability, $\hat{\bar{\omega}}(\bar{s}_t) = \max\big((\bar{\omega}(\bar{s}_t) - \eta_2 \boldsymbol{B}_{\bar{s}_t.x_t}, 0\big)$, $\Xi(\bar{s}_t) = p_{\bar{s}_t}(\bar{g}_t^{'}|x_t, \hat{\bar{\theta}}^{(j-1)})$ can be approximated as a Gaussian distribution by applying a Taylor expansion [8]. For the update equation of the AR coefficients $\bar{\alpha}_{\bar{s}}$, we can first estimate the autocorrelation sequence (15)

$$\bar{r}_{\bar{\alpha}_{\bar{s}}}^{(j)}[i] = \frac{\sum_r \sum_t \hat{\bar{\omega}}(\bar{s}_t) \bar{r}_{x_t}[i] \int (\bar{g}_t^{'})^{-1} \Xi(\bar{s}_t) d\bar{g}_t^{'}}{\sum_r \sum_t \hat{\bar{\omega}}(\bar{s}_t)}, \quad (15)$$

and then apply the Levinson-Durbin recursion algorithm [12]. Where $\bar{r}_{x_t}[i]$ denotes the autocorrelation sequence of $\boldsymbol{x}_t$.

The noise SARHMM can also be obtained by encouraging the sparsity to transition probabilities and observation probabilities. The noise SARHMM model is obtained using the standard Baum Welch algorithm [8] [11] using training data normalized by the long-term averaged noise gain. The noise gain variance $\ddot{\sigma}_{\ddot{s}}^2$ can be estimated as the sample variance of the logarithm of the excitation variances after the normalization, and the estimation process for the noise AR coefficients $\ddot{\alpha}_{\ddot{s}}$ are similar to speech AR coefficients $\bar{\alpha}_{\bar{s}}$.

4. SPEECH ENHANCEMENT USING SARHMMS

In addition to the ambiguity problem, existing ARHMM-based methods [4][8] have a second inherent problem: clearly audible noise remains between the harmonics of the estimated voiced speech. For this problem, we exploit the fact that the SARHMM provides estimates of both the noise and speech parameters, which aid in finding of a good noise model estimate. Therefore, instead of estimating speech directly, we first construct a noise estimator to get the noise spectrum. Then a Bayesian speech estimator is derived to obtain the enhanced speech.

4.1. Noise Estimation

In this section a noise estimator is constructed that is based on the SARHMM parameters of speech and noise. Following [4], we can obtain the noise spectrum estimation as

$$\hat{\ddot{\lambda}}(k) = \sum_{s_t} \frac{\omega(s_t)}{\Omega_t} \{[(1 - H_{s_t}(k)) Y(k)]^2 + H_{s_t}(k) \ddot{\lambda}_{\ddot{s}_t}(k)\}, \quad (16)$$

where t is the frame index, k is the frequency bin, $Y(k)$ is the kth spectral magnitude of noisy speech and $H_{s_t}(k)$ is the attenuation factor of the Wiener filter for state s_t, $\bar{\lambda}_{\bar{s}_t}(k)$ and $\ddot{\lambda}_{\ddot{s}_t}(k)$ are speech and noise power spectra associated to each composite state [13]. The $\omega(s_t)$ and Ω_t are defined by

$$\begin{aligned} \omega(s_t) &= \gamma(s_t) p_{s_t}(\boldsymbol{y}_t, \hat{\bar{g}}_t^{'}, \hat{\ddot{g}}_t^{'}), \\ \Omega_t &= \sum_{s_t} \gamma(s_t) p_{s_t}(\boldsymbol{y}_t, \hat{\bar{g}}_t^{'}, \hat{\ddot{g}}_t^{'}) = \sum_{s_t} \omega(s_t), \end{aligned} \quad (17)$$

where $\gamma(s_t)$ denotes the probability of being in the composite state s_t given all past noisy observation up to frame *t*-1, which is defined in [8].

4.2. Speech Estimation

Based on the estimated noise spectrum $\hat{\ddot{\lambda}}(k)$, we derive the Bayesian estimator to obtain the enhanced speech. For deriving the Bayesian estimator, we can minimize the expectation of a given cost function $d(X(k), \hat{X}(k))$ and obtain the speech estimate of $\hat{X}(k)$, where $X(k)$ is the kth spectral magnitude

of clean speech. The cost function $d(X(k), \hat{X}(k))$ that we minimize is [10]

$$d(X(k), \hat{X}(k)) = \frac{\left(X(k) - \hat{X}(k)\right)^2}{X(k)}, \quad (18)$$

Using a Gaussian statistical model [10], we obtain the speech estimate

$$\hat{X}(k) = \frac{\sqrt{v_k}\hat{\bar{\lambda}}(k)}{Y(k)} \frac{1}{\Gamma(1/2)\Phi(1/2, 1; -v_k)}, \quad (19)$$

where v_k is variable defined in [10], $\Gamma(\cdot)$ is the Gamma function, $\Phi(\cdot)$ is the confluent hyper-geometric function [10].

5. EXPERIMENTS AND RESULTS

In this section, the performance evaluation is discussed. Twenty four utterances (two female and two male speakers, each speaker for six sentences) were taken from the American sub-database of the NTT database. The sampling rate was 8 kHz. The White, Factory1, Factory2 and F16 noise from the NOISEX-92 [14] were used as the stationary noise and the Factory1, Factory2, F16 and Babble noise from the NOISEX-92 [14] were used to create three kinds of non-stationary noise: noise changing from Factory1 noise to F16 cockpit noise, noise changing from F16 noise to Babble noise, noise changing from Babble noise to Factory2 noise. The noise length was eight seconds and the noise changed abruptly after four seconds. The noisy utterances were created according to ITU-T P.56 standard [15] and the input SNRs of noisy speech are 0 dB, 5 dB, 10 dB and 15 dB, respectively.

The SARHMM used speech and noise states with a single mixture component, as more general mixtures lead to ambiguity. It operated with a frame size *N* is 256 samples. The samples were sine windowed with 50% overlap between adjacent frames. The regularization parameters were set to $p_1 = 0.4$ and $\eta_1 = 0.5$ and $p_2 = 0.4$ and $\eta_2 = 0.032$. The speech SARHMM had 64 states and an AR model of order 100. The one general noise SARHMM that covers many noise scenarios had 16 states and and an AR model of order six.

We compared the new method with, the MMSE method of Ephraim-Malah (MMSE) [3], MMSE estimator based on generalized gamma priors (GammaPrior) [16] and ARHMM-based gain modeling method (ARHMM) [8]. The segmental SNR (SNRseg) measure [17] and the perceptual evaluation of speech quality (PESQ) [18] were used to evaluate the performance. For the calculation of SNRseg, frames with an average energy 50 dB below the long-term average energy of the utterance were excluded.

The experimental results are presented in TABLE 1 and 2, respectively. From the two tables we can see that the proposed method can produce a higher average SNRseg improvement and a better speech quality. It is obvious that the proposed method performs better than the reference methods in non-stationary noise environments, while providing state-of-the-art performance for stationary conditions.

6. CONCLUSION

We introduced a sparse hidden Markov model (SARHMM)-based single-channel speech enhancement method. The method also includes an improved speech estimator that, in contrast to existing ARHMM and most codebook methods, enhances the spectral fine-structure of speech. Our results show that the method performs significantly better in non-stationary noise environments than reference enhancement procedures, while providing state-of-the-art performance for stationary conditions. The results confirm that sparsity, together with using only one mixture component for the observations, eliminates the ambiguity problem that ARHMM methods suffer from. It is better than existing ARHMM methods for stationary environments because of the improved speech estimator.

Table 1. Test Results of SNRseg Improvement

Methods	0dB	5dB	10dB	15dB
		Non-stationary		
MMSE	5.43	3.64	2.27	0.82
GammaPrior	6.13	5.19	4.28	3.33
ARHMM	5.50	3.70	1.76	0.17
SARHMM	7.69	6.16	5.13	4.06
		Stationary		
MMSE	6.45	4.56	2.79	0.81
GammaPrior	9.18	7.49	5.07	3.26
ARHMM	7.65	5.89	3.91	1.67
SARHMM	9.55	7.73	5.47	3.71

Table 2. Test Results of PESQ

Methods	0dB	5dB	10dB	15dB
		Non-stationary		
Noisy speech	1.654	1.954	2.286	2.614
MMSE	1.937	2.338	2.639	2.885
GammaPrior	2.052	2.408	2.754	2.985
ARHMM	1.846	2.244	2.575	2.856
SARHMM	2.169	2.501	2.814	3.080
		Stationary		
Noisy speech	1.565	1.839	2.158	2.496
MMSE	1.835	2.227	2.548	2.807
GammaPrior	2.155	2.514	2.816	3.067
ARHMM	1.807	2.209	2.501	2.753
SARHMM	2.228	2.572	2.864	3.091

7. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 61471014).

8. REFERENCES

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

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

[4] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal processing*, vol. 40, no. 4, pp. 725–735, 1992.

[5] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 163–176, 2006.

[6] ——, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 441–452, 2007.

[7] T. D. Christian, D. Sigg and J. M. Buhmann, "Speech enhancement using generative dictionary learning," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1698–1712, 2012.

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

[9] S. Bharadwaj, M. Hasegawa-Johnson, J. Ajmera, O. Deshmukh, and A. Verma, "Sparse hidden Markov models for purer clusters," in *Proc. IEEE Inf. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2013*, 2013, pp. 3098–3102.

[10] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 857–869, 2005.

[11] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden Markov models," *International Computer Science Institute*, vol. 4, no. 510, p. 126, 1998.

[12] B.-H. Juang and L. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 33, no. 6, pp. 1404–1413, 1985.

[13] 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, and Language Processing*, vol. 16, no. 4, pp. 835–846, 2008.

[14] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[15] ITU-T Rec, "Recommendation P.56-objective measurement of active speech level," *International Telecommunication Union, Geneva*, 1993.

[16] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1741–1752, 2007.

[17] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*. Prentice Hall Englewood Cliffs, NJ, 1988.

[18] ITU-T Rec, "Recommendation P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *International Telecommunication Union, Geneva*, 2001.