Speech Enhancement in Modulation Domain Using Codebook-based Speech and Noise Estimation

Vidhyasagar Mani, Benoit Champagne
Dept. of Electrical and Computer Engineering
McGill University, 3480 University St.
Montreal, Quebec, Canada, H3A 0E9
iitjsagar@gmail.com, benoit.champagne@mcgill.ca

Wei-Ping Zhu
Dept. of Electrical and Computer Engineering
Concordia University, 1455 Maisonneuve Blvd. West
Montreal, Quebec, Canada, H3G 1M8
weiping@ece.concordia.ca

***Abstract*—Conventional single-channel speech enhancement methods implement the analysis-modification-synthesis (AMS) framework in the acoustic frequency domain. In recent years, it has been shown that the extension of this framework to the modulation frequency domain may result in better noise suppression. However, this conclusion has been reached by relying on a minimum statistics approach for the required noise power spectral density (PSD) estimation, which is known to create a time frame lag when the noise is non-stationary. In this paper, to avoid this problem, we perform noise suppression in the modulation domain with speech and noise power spectra obtained from a codebook-based estimation approach. The PSD estimates derived from the codebook approach are used to obtain a minimum mean square error (MMSE) estimate of the clean speech *modulation* magnitude spectrum, which is combined with the phase spectrum of the noisy speech to recover the enhanced speech signal. Results of objective evaluations indicate improvement in noise suppression with the proposed codebook-based speech enhancement approach, particularly in cases of non-stationary noise.**[1]

***Index Terms*—Speech enhancement, modulation domain, MMSE estimation, LPC codebooks**

I. Introduction

Speech enhancement involves the suppression of background noise from a desired speech signal while ensuring that the incurred distortion is within a tolerable limit. Some of the most commonly used single channel speech enhancement methods include spectral subtraction [1], [2], Wiener filtering [3], and MMSE short-time spectral amplitude (STSA) estimation [4], [5]. These methods typically involve implementation of the following three-stage framework known as AMS [6], [7]: (1) Analysis, in which the short-time fourier transform (STFT) is applied on successive frames of the noisy speech signal; (2) Modification, where the spectrum of the noisy speech is altered for achieving noise suppression, and; (3) Synthesis, where the enhanced speech is recovered via inverse STFT and overlap-add (OLA) synthesis.

In past years, research has shown that extension of this framework into the modulation domain may result in improved noise suppression and better speech quality [8], [9]. For instance, in the case of spectral subtraction, musical noise distortion is lesser when the subtraction is performed in the modulation domain than in the conventional frequency domain [8]. Extension of the MMSE-STSA estimator to the modulation domain, in the form of the modulation magnitude estimator (MME) [9], has also shown positive results. The interest towards this framework extension is further motivated by physiological evidence [10]–[12], which underlines the significance of modulation domain information in speech analysis.

Most speech enhancement algorithms, including those operating in the modulation domain, require an estimate of the background noise PSD which is typically obtained via a minimum statistics [13] approach. Minimum statistics and its offshoots [14], [15] assume that the background noise exhibits a semi-stationary behaviour (i.e. slowly changing statistics) while performing its estimation. This may not be the case in acoustic environments with rapidly changing background, e.g., a street intersection with passing vehicles or a busy airport terminal. In such cases, the noise PSD cannot be tracked properly and speech enhancement algorithms may perform poorly.

Codebook based approaches [16]–[20], which fit under the general category of unsupervised learning [21], try to overcome this limitation by estimating the noise parameters based on *a priori* knowledge about different speech and noise types. In these approaches, joint estimation of the speech and noise PSD is performed on a frame-by-frame basis by exploiting *a priori* information stored in the form of trained codebooks of short-time parameter vectors. Examples of such parameters include gain normalized linear predictive (LP) coefficients [16]–[19] and cepstral coefficients [20].

The use of these codebook methods in the acoustic AMS framework has shown promising results in the enhancement of speech corrupted by non-stationary noise. However, to the best of our knowledge, they have not been applied yet to the modulation domain framework. In this work, we conjecture that codebook methods can indeed bring similar benefits to the enhancement of noisy speech in the modulation domain by providing more accurate estimation of the noise PSD in non-stationary environments, and validate this hypothesis experimentally.

Specifically, the new speech enhancement method that we propose in this paper incorporates codebook assisted noise and speech PSD estimation into the modulation domain framework. We use codebooks of linear prediction coefficients and gains obtained by training with the Linde-Buzo-Gray (LBG) algorithm [22]. The PSD estimates derived from the codebook approach are used to calculate a gain function based on the MMSE criterion [9], which is applied to the modulation magnitude spectrum of the noisy speech in order to suppress noise. Results of objective evaluations indicate improvement in noise suppression with the proposed codebook-based speech enhancement method, especially in cases of non-stationary noise.

II. Acoustic versus modulation domain processing

A. AMS in the Acoustic Frequency Domain

Conventional speech enhancement methods implement the AMS framework in the acoustic frequency domain, where the acoustic frequency spectrum of a speech signal is defined by its STFT. To

[1] Funding for this work was provided by a CRD grant from the Natural Sciences and Engineering Research Council of Canada under sponsoring from Microsemi Corporation (Ottawa, Canada).

978-1-4799-7591-4/15/$31.00 ©2015 IEEE

this end, an additive noise model is assumed, i.e.,

$$x[n] = s[n] + d[n], \tag{1}$$

where $x[n]$, $s[n]$ and $d[n]$ refer to the noisy speech, clean speech and noise signals respectively, while $n \in \mathbb{Z}$ is the discrete-time index. STFT analysis of (1) results in,

$$X(\nu, k) = S(\nu, k) + D(\nu, k) \tag{2}$$

where $X(\nu, k)$, $S(\nu, k)$ and $D(\nu, k)$ refer to the STFTs of the noisy speech, clean speech and noise signals, respectively, and where k is the discrete acoustic frequency index. The STFT $X(\nu, k)$ is obtained from,

$$X(\nu, k) = \sum_{l=-\infty}^{\infty} x(l) w(\nu F - l) e^{-2jkl\pi/N} \tag{3}$$

where $w(l)$ is a windowing function of duration N samples, and F is the frame advance. In this work, the Hamming window is used for this purpose [7]. The STFT of a signal is represented by its acoustic magnitude and phase spectra as,

$$X(\nu, k) = |X(\nu, k)| e^{j\angle X(\nu,k)} \tag{4}$$

Speech enhancement methods, such as spectral subtraction [1] or MMSE-STSA [4], implement the modification part of the AMS framework by modifying the noisy magnitude spectrum whilst retaining the phase spectrum. Synthesis of the enhanced signal is performed by inverse STFT followed by OLA synthesis.

B. Modulation Domain Enhancement

The calculation of the short time modulation spectrum involves performing STFT analysis on time trajectories of the individual acoustic frequency components of the signal STFT. The magnitude spectrum of the noisy speech in each acoustic frequency bin, i.e. $|X(\nu, k)|$, is first windowed and then Fourier transformed again, resulting into,

$$Z(t, k, m) = \sum_{\nu=-\infty}^{\infty} |X(\nu, k)| w_M(tF_M - \nu) e^{-2j\nu m\pi/M} \tag{5}$$

where $w_M(\nu)$ is the so-called modulation window of length N_M, $m \in \{0, ..., M-1\}$ is the modulation frequency index, t is the modulation time-frame index, and F_M is the frame advance in the modulation domain. The resulting modulation spectrum can be expressed in polar form as,

$$Z(t, k, m) = |Z(t, k, m)| e^{j\angle Z(t,k,m)} \tag{6}$$

where $|Z(t, k, m)|$ is the *modulation* magnitude spectrum and $\angle Z(t, k, m)$ is the modulation phase spectrum.

Speech enhancement in the modulation domain involves spectral modification of the modulation magnitude spectrum while retaining the phase spectrum,

$$\hat{S}(t, k, m) = G(t, k, m) Z(t, k, m) \tag{7}$$

where $G(t, k, m) > 0$ is a processing gain. Following this operation, the enhanced time-domain signal is recovered by applying inverse STFT and OLA operations twice. Previous works [8], [9] suggest that enhancement approaches applied in the modulation domain perform better than their traditional acoustic domain counterparts. In this work, the MMSE estimator of the modulation magnitude spectrum, also known as MME [9], will be used as a basis for developing the proposed codebook-based speech enhancement method.

III. Codebook-based speech and noise estimation

A. Overview

Various noise estimation algorithms are available in the literature to estimate the background noise PSD, needed to perform noise suppression in speech enhancement. In algorithms based on minimum statistics [13], [14], which are widely applied, the noise PSD is updated by tracking the minima of a smoothed version of $|X(\nu, k)|^2$ within a finite window. Tracking the minimum power in this way results in a frame lag in the estimated PSD. This lag can lead to highly inaccurate results in the case of non-stationary noise. The basis for the codebook-based speech and noise PSD estimation approach in [17]–[20] is the observation that the spectra of speech and different noise classes can be approximately described by few representative models' spectra. These spectra are stored in finite codebooks as quantized vectors of short-time parameters (e.g., LP coefficients) and serve as the *a priori* knowledge of the respective signals. The use of *a priori* information about noise eliminates the dependence on buffers of past data. This makes the estimation robust to spectral variations in non-stationary noise conditions [16].

B. PSD Model

For the additive noise model (1), under the assumption of uncorrelated speech and noise signals, the PSD of the noisy speech can be represented as,

$$P_{xx}(\omega) = P_{ss}(\omega) + P_{dd}(\omega), \quad \omega \in [0, 2\pi) \tag{8}$$

where $P_{ss}(\omega)$ and $P_{dd}(\omega)$ are the clean speech and background noise PSD, respectively, and $\omega \in [0, 2\pi)$ is the normalized angular frequency. The PSD shape of signal $y[n]$, where $y \in \{s, d\}$ stands for either the speech or noise, can be modelled in terms of its LP coefficients and corresponding excitation variance as,

$$P_{yy}(\omega) = g_y \, \overline{P}_{yy}(\omega) \tag{9}$$

where $\overline{P}_{yy}(\omega)$ is the gain normalized spectral envelope and g_y is the excitation gain (or variance). The former is given by,

$$\overline{P}_{yy}(\omega) = \left|1 + \sum_{k=1}^{p} a_k^y e^{j\omega k}\right|^{-2} \tag{10}$$

where $\{a_k^y\}_{k=1}^p$ are the LP coefficients, represented here by vector $\boldsymbol{\theta}_y = [a_1^y, ..., a_p^y]$, and p is the model order chosen.

C. Codebook Generation

In this work, two different codebooks of short-time spectral parameters, one for the speech and the other for the noise, are generated from training data comprised of multiple speaker signals and different noise types. The codebook generation comprises the following steps: segmentation of the training speech and noise data into frames with 20-40ms duration; computation of LP coefficients $\{a_k^y\}_{k=1}^p$ for each frame; vector quantization of the LP coefficient vectors $\boldsymbol{\theta}_y$ using the LBG algorithm to obtain the required codebook [22]. The LBG algorithm forms a set of median cluster vectors which best represent the given input set of LP coefficient vectors. Optimal values have to be chosen empirically for the size of the speech and noise codebooks, considering the trade-off between PSD estimation accuracy and complexity. In the sequel, we shall represent the speech and noise codebooks so obtained as $\{\boldsymbol{\theta}_s^i\}_{i=1}^{N_s}$ and $\{\boldsymbol{\theta}_d^j\}_{j=1}^{N_d}$, where vectors $\boldsymbol{\theta}_s^i$ and $\boldsymbol{\theta}_d^j$ are the corresponding i-th and j-th codebook entries, and N_s and N_d are the codebook sizes, respectively.

In addition to the codebook vectors generated from training on noise data, during the estimation phase, the noise codebook is supplemented by one extra vector. The latter is updated for every frame based on a noise PSD estimate obtained using a MS method [13], [14]. This provides robustness in dealing with noise types which may not be present in the training set.

D. Gain Adaptation

Each codebeook entry, i.e., $\boldsymbol{\theta}^i_s$ or $\boldsymbol{\theta}^j_d$, can be used to compute a corresponding gain normalized spectral envelope, respectively $\overline{P}^i_{ss}(\omega)$ or $\overline{P}^j_{dd}(\omega)$ by means of relations (10). To obtain the final PSD shape as in (9), however, the resulting envelope needs to be scaled by a corresponding excitation gain, which we denote as g^i_s and g^j_d, respectively. In this work, we use an adaptive approach whereby the excitation gains for the speech and noise codebooks are updated every frame based on the observed noisy speech magnitudes $|X(\nu, k)|$.

Specifically, for every possible combination of vectors $\boldsymbol{\theta}^i_s$ and $\boldsymbol{\theta}^j_d$ from the speech and noise codebooks, respectively, the corresponding gains g^i_s and g^j_d at the ν-th frame are obtained by minimizing the Itakura-Saito distance measure between an estimated PSD and the squared magnitude spectrum $|X(\nu, k)|^2$ of the noisy speech over the frequency domain. In this calculation, the estimated PSD is defined as the sum of the gain-adapted speech and noise envelopes, i.e.,

$$P^{ij}_{xx} = g^i_s \overline{P}^i_{ss}(\omega) + g^j_d \overline{P}^j_{dd}(\omega). \tag{11}$$

The final optimum values of g^i_s and g^j_d, which can be interpreted as conditional ML estimates, are approximated as in [18].

E. Joint PSD Estimation

The joint estimation of the speech and noise PSD is done on a frame by frame basis. Let $\boldsymbol{\theta} = [\boldsymbol{\theta}_s, \boldsymbol{\theta}_d, g_s, g_d]$ denote the vector of unknown parameters to be estimated, and from which speech and noise PSD can be determined through (9)-(10). Following [19], we adopt an MMSE framework for the estimation of parameter vector $\boldsymbol{\theta}$. This framework makes it possible to simultaneously estimate the LP coefficients (and excitation gains) of two linear processes that additively overlap with each other.

To this end, the noisy speech signal $x[n]$ in (1) is assumed to follow a multivariate normal distribution when conditioned on $\boldsymbol{\theta}$,

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{1}{(2\pi)^{N/2}\det(\boldsymbol{R}_{xx})^{1/2}} e^{-(1/2)(\boldsymbol{x}^T \boldsymbol{R}_{xx}^{-1} \boldsymbol{x})} \tag{12}$$

where $\boldsymbol{x} = [x[\nu F+1], \ldots, x[\nu F+N]]^T$ is the observed data vector at the ν-th frame and $\boldsymbol{R}_{xx} = E\{\boldsymbol{x}\boldsymbol{x}^T\}$ is the associated covariance matrix. Under the previous modeling assumptions, the latter can be written as the sum of the speech and noise covariance matrices, i.e., $\boldsymbol{R}_{xx} = \boldsymbol{R}_{ss} + \boldsymbol{R}_{dd}$. In turn, $\boldsymbol{R}_{ss}$ and $\boldsymbol{R}_{dd}$ are functions of the corresponding LP coefficients and excitation gains, as in $\boldsymbol{R}_{ss} = g_s(\boldsymbol{A}_s^T \boldsymbol{A}_s)^{-1}$ where $\boldsymbol{A}_s$ is an $N \times N$ Toeplitz lower triangular matrix derived from $\boldsymbol{\theta}_s^T$.

The equation for the conditional distribution $p(\boldsymbol{x}|\boldsymbol{\theta})$ in (12) involves a matrix inversion, which is computationally expensive. For a simpler and less time consuming computation, the covariance matrices $\boldsymbol{R}_{ss}$ and $\boldsymbol{R}_{dd}$ can be approximated as circulant matrices [17], thereby reducing (12) to,

$$\begin{aligned} \ln p(\boldsymbol{x}|\boldsymbol{\theta}) \approx & -\frac{N}{2}\ln 2\pi - \frac{1}{2}\sum_{k=0}^{N-1} \ln(g_s \overline{P}_{ss}(\omega_k) + g_d \overline{P}_{dd}(\omega_k)) \\ & - \frac{1}{2}\sum_{k=0}^{N-1} \frac{|X(\nu, \omega_k)|^2}{g_s \overline{P}_{ss}(\omega_k) + g_d \overline{P}_{dd}(\omega_k)} \end{aligned} \tag{13}$$

where $\omega_k = \frac{2\pi k}{N}$. Equation (13) is a reasonable approximation of (12) for large frame sizes N.

With the help of the estimated excitation gains at the ν-th frame, we can define for each pair of speech and noise codebook vectors $\boldsymbol{\theta}^i_s$ and $\boldsymbol{\theta}^j_d$ a complete codebook-based parameter vector $\boldsymbol{\theta}^{ij} = [\boldsymbol{\theta}^i_s, \boldsymbol{\theta}^j_d, g^i_s, g^j_d]$. The joint MMSE estimation of the unknown parameter vector $\boldsymbol{\theta}$ is implemented by carrying numerical integration over the product codebook of vectors $\boldsymbol{\theta}^{ij}$ so obtained, as given by [19]:

$$\hat{\boldsymbol{\theta}}_{\text{MMSE}} \approx \frac{1}{N_s N_d} \sum_{i=1}^{N_s} \sum_{j=1}^{N_d} \boldsymbol{\theta}^{ij} \frac{p(\boldsymbol{x}|\boldsymbol{\theta}^{ij})}{p(\boldsymbol{x})} \tag{14}$$

$$p(\boldsymbol{x}) \approx \frac{1}{N_s N_d} \sum_{i=1}^{N_s} \sum_{j=1}^{N_d} p(\boldsymbol{x}|\boldsymbol{\theta}^{ij}). \tag{15}$$

These equations provide a fair approximation to the MMSE estimate under the assumptions that the codebook is sufficiently large and the unknown parameter vector $\boldsymbol{\theta}$ is uniformly distributed.

IV. Incorporation of Codebook-based PSD into the Modulation Magnitude Estimator

The MME method [9] is an extension of the widely used acoustic domain based MMSE spectral amplitude estimator [4], into the modulation domain. In the MME method, the clean speech modulation magnitude spectrum is estimated from the noisy speech by minimizing the mean square error, denoted as $\mathcal{E}$, between the clean and estimated speech, i.e.,

$$\mathcal{E} = E[(|S(t,k,m)| - |\hat{S}(t,k,m)|)^2] \tag{16}$$

where $|S(t,k,m)|$and$|\hat{S}(t,k,m)|$ denote the modulation magnitude spectra of the clean and estimated speech, respectively. Using this MMSE criterion, the modulation magnitude spectrum of the clean speech can be estimated from the noisy speech as,

$$|\hat{S}(t,k,m)| = G(t,k,m)|Z(t,k,m)| \tag{17}$$

where $G(t,k,m)$ is the MME spectral gain function and $Z(t,k,m)$ is the modulation spectrum of the noisy speech from (5). The MME gain function is given by [9],

$$G(t,k,m) = \frac{\sqrt{\pi\nu}}{2\gamma} \exp\left(\frac{-\nu}{2}\right) \left[(1+\nu) I_0\left(\frac{-\nu}{2}\right) + \nu I_1\left(\frac{-\nu}{2}\right)\right] \tag{18}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified bessel functions of order zero and one, respectively, and the parameter $\nu \equiv \nu(t,k,m) = \frac{\xi}{1+\xi}\gamma$ is defined in terms of the *a priori* and *a posteriori* SNRs ξ and γ.

It is precisely in the calculation of these SNR parameters that we make use of the codebook-based PSD estimates. In this work, the *a posteriori* SNR is estimated as,

$$\hat{\gamma}(t,k,m) = \frac{|Z(t,k,m)|^2}{|\hat{D}(t,k,m)|^2} \tag{19}$$

where $|\hat{D}(t,k,m)|^2$ is an estimate of the noise power in the modulation domain. This quantity is obtained by applying the STFT (over frame index ν) to the square-root of the codebook-based noise PSD estimate, and then squaring the result. Specifically,

$$\hat{D}(t,k,m) = \sum_{\nu} \sqrt{P_{dd}(\nu,k)}\, w_M(tF_M - \nu) e^{-2j\nu m\pi/M} \tag{20}$$

where $P_d(\nu,k)$ is the noise PSD estimate obtained at the ν-th frame through codebook-based MMSE estimation.

To reduce spectral distortion the following "decision directed" approach is employed to obtain the value of the *a priori* SNR,

$$\hat{\xi}(t,k,m) = \alpha \frac{|\hat{S}(t-1,k,m)|^2}{|\hat{D}(t-1,k,m)|^2} + (1-\alpha)\frac{|C(t,k,m)|^2}{|\hat{D}(t,k,m)|^2} \quad (21)$$

where $|C(t,k,m)|^2$ is an estimate of the clean speech power in the modulation domain and $0 < \alpha < 1$ is a control factor which acts as a trade-off between noise reduction and speech distortion. Similar to (20), $C(t,k,m)$ is obtained by applying the STFT to the square-root of $P_{ss}(\nu,k)$, i.e. the codebook-based PSD estimate of the clean speech at the ν-th frame.

The estimated modulation magnitude spectrum, $|\hat{S}(t,k,m)|$ in (15), is transformed to the acoustic frequency domain by applying inverse STFT followed by OLA synthesis. The resulting spectrum is combined with the phase spectrum of the noisy speech to obtain the enhanced speech spectrum. The latter is mapped back to the time by performing inverse STFT followed by OLA synthesis.

V. Experimental Evaluation

In this section we describe objective evaluation experiments that were performed to assess the performance of the proposed algorithm, referred to as codebook-based MME (CB-MME). Other enhancement methods, including the acoustic domain MMSE-STSA [4] and modulation domain MME [9], were also evaluated for comparison.

A. Methodology

Speech utterances of two male and two female speakers from the TSP [23] and TIMIT databases were used for conducting the experiments, along with different types of noise samples from the NoiseX92 [24] and Sound Jay [25] databases, including babble, street and restaurant noise. In addition, a non-stationary (i.e. amplitude modulated) Gaussian white noise was also considered. All the speech and noise files were uniformly sampled at a rate of 16kHz. The LP coefficient order p was set to 10 for both speech and noise codebooks. A 7-bit speech codebook was trained with 7.5 minutes of clean speech from the above mentioned sources. (i.e 55 short sentences for each speaker). A 4-bit noise codebook was trained using over 1 minute of noise data from the available databases (i.e. about 15s for each noise type). For the testing, i.e. objective evaluation of the various algorithms, noisy speech files were generated by adding scaled segments of noise to the clean speech. For each speaker, 3 sentences were selected and combined with the four different types of noise, properly scaled to obtain the desired SNR values of 0 and 5dB. The speech and noise samples used for testing were different from those used to train the two codebooks.

Fine tuning of parameters is crucial for the performance of the proposed enhancement method. The acoustic frame duration was chosen to be 32ms, while the values of the other analysis parameters where chosen empirically as follows: acoustic frame advance F = 4ms, modulation frame duration N_M = 80, modulation frame advance F_M = 8ms and control factor α = 0.95.

For the objective evaluation of the enhanced speech, we used the perceptual evaluation of speech quality (PESQ) and the segmental SNR (SegSNR) as performance measures. PESQ [26] is widely used for automated assessment of speech quality as experienced by a listener, where higher PESQ values indicate a better speech quality. SegSNR is defined as the average SNR calculated over short segments of speech; higher SegSNR values indicate lesser background noise.

TABLE I: PESQ values

Input	SNR	Noisy	MMSE	MME	CB-MME
NS-white	0 dB	1.75	1.78	2.04	**2.24**
	5 dB	2.06	2.19	2.46	**2.58**
Street	0 dB	1.72	1.85	1.95	**2.07**
	5 dB	2.01	2.17	2.30	**2.40**
Restaurant	0 dB	1.78	1.84	1.87	**2.04**
	5 dB	2.13	2.20	2.27	**2.37**
Babble	0 dB	1.67	1.83	1.93	**2.07**
	5 dB	2.04	2.19	2.30	**2.43**

TABLE II: Segmental SNR values (dB)

Input	SNR	Noisy	MMSE	MME	CB-MME
NS-white	0 dB	-2.02	-1.19	0.57	**1.63**
	5 dB	1.55	2.60	3.75	**5.04**
Street	0 dB	-2.75	-0.96	0.47	**1.09**
	5 dB	0.72	1.35	1.91	**2.94**
Restaurant	0 dB	-2.44	-2.31	-0.59	**0.71**
	5 dB	1.14	1.43	2.07	**3.67**
Babble	0 dB	-3.02	-2.24	-0.85	**0.47**
	5 dB	0.84	1.28	2.36	**3.16**

B. Results & Discussion

The PESQ and SegSNR results for different noises at SNR of 0 and 5dB are reported in Tables I and II, respectively. It can be seen that the proposed CB-MME method performs better than the MME and MMSE methods, for both performance metrics under consideration. Results for other SNR and noise types (not shown) show a similar trend. Informal listening tests concur with the objective results. The proposed CB-MME method seems to suppress non-stationary elements of background noise better than MMSE and MME, at the expense of some slight distortion in the enhanced speech. This is mainly due to the use of a codebook-based approach, which performs on-line noise PSD estimation on a frame-by-frame basis based on current observation, as opposed to the MS approach used in the MMSE and MME algorithms, which relies on a long buffer of past frames. The slight distortion could be caused by the spectral mismatch between the codebook-based speech PSD estimate and the actual one, which remains a topic for future study.

VI. Conclusion

In this paper, we have proposed a new speech enhancement method that performs noise suppression in the modulation domain with speech and noise PSD obtained from a codebook-based estimation approach. We use codebooks of linear prediction coefficients and gains obtained by training with the LBG algorithm. The PSD estimates derived from the codebooks were used to calculate an MMSE gain function, which was applied to the modulation magnitude spectrum of the noisy speech in order to suppress noise. Results of objective evaluation showed improvements in the suppression of non-stationary noise with the proposed CB-MME approach.

References

[1] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, pp. 113-120, Apr. 1979.

[2] N. Virag, “Single channel speech enhancement based on masking properties of the human auditory system,” *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 126-137, Mar. 1999.

[3] J. Chen, J. Benesty, Y. Huang, “New insights into the noise reduction Wiener filter,” *IEEE Trans. Acoust. Speech Signal Process.*, vol. 14, pp. 1218-1234, Jul. 2006.

[4] Y. Ephraim, D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust. Speech Signal Process.*, vol. 32, pp. 1109-1121, Dec. 1984.

[5] E. Plourde, B. Champagne, “Generalized Bayesian estimators of the spectral amplitude for speech enhancement,” *IEEE Signal Process. Letters*, vol. 16, pp. 485-488, Jun. 2009.

[6] D. Griffin, J. Lim, “Signal estimation from modified short-time Fourier transform,” *IEEE Trans. Acoust. Speech Signal Process.*, vol. 2, pp. 236-243, Apr. 1984.

[7] T. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Prentice Hall, 2002.

[8] K. Paliwal, K. Wojcicki, B. Schwerin, “Single-channel speech enhancement using spectral subtraction in the short-time modulation domain,” *Speech Commun.*, vol. 52, no. 5, pp. 450-475, May 2010.

[9] K. Paliwal, B. Schwerin, K. Wojcicki, “Speech enhancement using minimum mean-square error short-time spectral modulation magnitude estimator,” *Speech Commun.*, vol. 54, no. 2, pp. 282-305, Feb. 2012.

[10] L. Atlas, S. Shamma, “Joint acoustic and modulation frequency,” *EURASIP J. on Applied Signal Process.*, vol. 7 , pp. 668-675, Jan. 2003.

[11] A. I. Shim, B. G. Berg, “Estimating critical bandwidths of temporal sensitivity to low-frequency amplitude modulation,” *J. Acoustical Society of America,* vol. 5, pp. 2834-2838, May 2013.

[12] K. Paliwal, B. Schwerin, “Modulation Processing for Speech Enhancement,” Chap. 10 in T. Ogunfunmi, R. Togneri and M. Narasimha, Eds., *Speech and Audio Processing for Coding, Enhancement and Recognition,* Springer 2015.

[13] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504-512, Jul. 2001.

[14] I. Cohen, “Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging ,” *IEEE Trans. on Speech and Audio Process.*, vol. 11, pp. 466-475, Sep. 2003.

[15] V. Stahl, A. Fischer, R. Bippus, “Quantile based noise estimation for spectral subtraction and wiener filtering,” *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Process.*, vol.3, pp. 1875-1878, Jun. 2000.

[16] S. Srinivasan, J. Samuelsson, W. B. Kleijn, “Speech enhancement using a-priori information,” *Proc. Eurospeech,*, pp. 1405-1408, Sep. 2003.

[17] M. Kuropatwinski, W. B. Kleijn, “Estimation of the short-term predictor parameters of speech under noisy conditions,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1645-1655, Sep. 2006.

[18] S. Srinivasan, J. Samuelsson, W. B. Kleijn, “Codebook driven short-term predictor parameter estimation for speech enhancement,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 1, pp. 163-176, Jan. 2006.

[19] S. Srinivasan, J. Samuelsson, W. B. Kleijn, “Codebook-based Bayesian speech enhancement for nonstationary environments,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 441-452, Feb. 2007.

[20] T. Rosenkranz, “Modeling the temporal evolution of LPC parameters for codebook-based speech enhancement,” *Int. Symp. on Image and Signal Process. and Analysis, Salzburg*, pp. 455-460 , Sep. 2009.

[21] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning, 2nd Ed.* Springer, 2009.

[22] Y. Linde, A. Buzo, R. M. Gray, “An algorithm for vector quantizer design,” *IEEE Trans. Communications,* vol. 28, no. 1, pp. 84-95, Jan. 1980.

[23] P. Kabal, McGill University, “TSP speech database,” Tech. Rep., 2002.

[24] Rice University, “Signal processing information base: noise data.” Available online: http://spib.rice.edu/spib/select _noise.html.

[25] Sound Jay, “Ambient and special sound effects.” Available online: http://www.soundjay.com/ambient-sounds-2.html.

[26] ITU-T. P.862, “Perceptual evaluation of speech quality (PESQ): and objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” Tech. Rep., 2000.

[27] E. Vincent, R. Gribonval, C. Fevotte, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio, Speech and Language Process.*, vol. 14, no. 4, pp. 1462-1469, Jul. 2006.