EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Robust Bayesian estimation for context-based speech enhancement

Devireddy Hanumantha Rao Naidu[1]* and Sriram Srinivasan[2]

Abstract

Model-based speech enhancement algorithms that employ trained models, such as codebooks, hidden Markov models, Gaussian mixture models, etc., containing representations of speech such as linear predictive coefficients, mel-frequency cepstrum coefficients, etc., have been found to be successful in enhancing noisy speech corrupted by nonstationary noise. However, these models are typically trained on speech data from multiple speakers under controlled acoustic conditions. In this paper, we introduce the notion of context-dependent models that are trained on speech data with one or more aspects of context, such as speaker, acoustic environment, speaking style, etc. In scenarios where the modeled and observed contexts match, context-dependent models can be expected to result in better performance, whereas context-independent models are preferred otherwise. In this paper, we present a Bayesian framework that automatically provides the benefits of both models under varying contexts. As several aspects of the context remain constant over an extended period during usage, a memory-based approach that exploits information from past data is employed. We use a codebook-based speech enhancement technique that employs trained models of speech and noise linear predictive coefficients as an example model-based approach. Using speaker, acoustic environment, and speaking style as aspects of context, we demonstrate the robustness of the proposed framework for different context scenarios, input signal-to-noise ratios, and number of contexts modeled.

Keywords: Bayesian; Codebook; Context; Noise reduction; Speech enhancement

1 Introduction

Speech enhancement pertains to the processing of speech corrupted by noise, echo, reverberation, etc. to improve its quality and intelligibility. In this paper, by speech enhancement, we refer to the problem of noise reduction. It is relevant in several scenarios, for example, mobile telephony in noisy environments, such as restaurants and busy traffic, suffers from unclear communication. Also, speech recognition units [1] and hearing aids [2] require speech enhancement as a preprocessing algorithm.

Speech enhancement algorithms can be broadly classified into single- and multi-channel algorithms based on the number of microphones used to acquire the input noisy speech. Multi-channel algorithms exhibit superior performance because of the additional spatial information available about the noise and speech sources. However, the need for single-channel speech enhancement cannot be ignored. For example, single microphone systems are preferred in low-cost mobile units. In addition, multi-channel methods include a single-channel algorithm as a post-processing step to suppress diffuse noise. In this paper, we focus on single-channel speech enhancement.

Single-channel speech enhancement has been a challenging research problem for the last four decades. Several techniques have been devised to arrive at efficient solutions for the problem. Among these, spectral subtraction is one of the earliest and simplest techniques [3]. Herein, an estimate of the noise magnitude spectrum is subtracted from the observed noisy magnitude spectrum to obtain an estimate of the clean speech magnitude spectrum. Several variations of this technique have been developed over the years [4-7]. Methods based on a statistical model of speech to estimate the speech spectral amplitude such as the minimum mean square error short-time spectral amplitude estimator (MMSE-STSA) method have been found to be successful [8-10]. The statistical approach explicitly uses the probability density function (pdf) of the speech and noise DFT coefficients. Also, it

*Correspondence: dhanumantharao@sssihl.edu.in
[1] Department of Mathematics and Computer Science, Sri Sathya Sai Institute of Higher Learning, Prasanthi Nilayam, Anantapur, Andhra Pradesh 515134, India
Full list of author information is available at the end of the article

allows consideration of non-Gaussian prior distributions [11] and different ways of modeling the spectral data [12,13]. Subspace-based algorithms [14] assume the clean speech to be confined to a subspace of the noisy space. The noisy vector space is decomposed into noise-only and speech-plus-noise subspaces. The noise subspace components are suppressed, and the speech-plus-noise subspace components are further processed. A comprehensive survey of these techniques is provided in [15]. However, most of these methods depend on an accurate estimate of the noise power spectrum, for example, estimation of the noise magnitude spectrum during silent segments in [3], or *a priori* signal-to-noise ratio (SNR) estimation in [9], or estimation of the noise covariance matrix in the subspace-based methods.

Noise estimation algorithms mainly include voice activity detector (VAD) [16,17] and buffer-based methods [18-20]. While VADs are unreliable at low SNRs, the buffer-based methods are not fast enough to track the quickly varying noise in nonstationary noise conditions. Thus, while these algorithms perform well in stationary noise, their accuracy deteriorates under nonstationary conditions. An improvement over these algorithms is provided in [21] wherein a recursive approach is employed for online noise power spectral density (PSD) tracking by analytically retrieving the prior and posterior probabilities of speech absence, and noise statistics, using a maximum likelihood-based criterion. A low-complexity, fast noise tracking algorithm is proposed in [22,23].

Speech enhancement algorithms which employ trained models, such as codebooks [24-28], hidden Markov models (HMM) [29-31], Gaussian mixture models (GMM) [32], non-negative matrix factorization (NMF) models [33], dictionaries [34], etc., for speech and noise data are able to process noisy speech with sufficient accuracy even under nonstationary noise conditions. For example, codebook-based speech enhancement (CBSE) algorithms [25,26] estimate the noise power spectrum for short segments of noisy speech, thus tracking nonstationary noise better than the buffer-based methods [18]. However, model-based methods typically employ *a priori* speech models which are trained on speech data from multiple speakers. For applications where the input noisy speech is more frequent from a particular speaker, such as in mobile telephony, it is desirable to exploit the speaker dependency for better speech enhancement. Similarly, it might be beneficial to consider models trained on or adapted to a specific acoustic environment or language. In this paper, we introduce the notion of context-dependent (CD) models, where by the word 'context', we refer to one or more aspects such as the speaker, acoustic environment, emotion, language, speaking style, etc. of the input noisy speech. By employing CD models, improved enhancement of noisy speech can be expected. These models can be adapted online from a context-independent (CI) model during high SNR regions of the input signal. In this paper, we assume the availability of such adapted CD models and focus on the enhancement using the converged models.

When the context of the noisy input matches the context of the data used to train the model, CD models are expected to result in better speech enhancement than CI models. We refer to such scenarios as context match scenarios. However, in practice, the modeled and observed contexts may not always match, leading to a context mismatch. In such scenarios, a CD model may lead to poorer results, and so the CI model would be preferred. Thus, what is required is a method that retains the benefits of both the CD and CI models and provides robust results irrespective of the scenario at hand.

In this paper, we introduce a Bayesian framework to optimally combine the estimates from the CD and CI models to achieve robust speech enhancement under varying contexts. As different aspects of context can be expected to remain constant for an extended duration in the input noisy signal, the framework considers past information to improve the estimation process. Also, in practice, different aspects of context may occur at the same time. So, the framework is designed to include several codebooks at the same time.

As an example of the model-based algorithm, we use the CBSE technique that employs trained models of speech and noise linear predictive (LP) coefficients as priors [26]. A part of this work has been presented in [35]. This papers extends [35] by incorporating memory-based estimation, considers the use of multiple CD models, and presents a detailed experimental analysis for different noise types, input SNRs, and aspects of context. The framework developed is general and can be used for other representations such as mel-frequency cepstrum coefficients, higher resolution PSDs, as well as other models such as GMMs, HMMs, and NMF.

The remainder of the paper is organized as follows. In the next section, a brief outline of the CBSE techniques [25,26] is provided. Following this, we derive the memory-based Bayesian framework to optimally combine estimates from several codebooks (CD/CI). Thereafter, we present the experimental results for the proposed framework under varying contexts, noise types, and input SNRs. Finally, we summarize the conclusions.

2 Codebook-based speech enhancement

Consider an additive noise model of the observed noisy speech $y(n)$:

$$y(n) = x(n) + w(n), \tag{1}$$

where n is the time index, $x(n)$ is the clean speech signal, and $w(n)$ is the noise signal.

We assume that speech and noise are statistically independent and follow zero-mean Gaussian distribution. Under these assumptions, Equation 1 leads to the following relation in the frequency domain:

$$P_y(\omega) = P_x(\omega) + P_w(\omega), \tag{2}$$

where $P_y(\omega)$, $P_x(\omega)$, and $P_w(\omega)$ are PSDs of the observed noisy speech, clean speech, and noise respectively, and ω is the angular frequency.

Consider a short-time segment of the observed noisy speech given by a vector $\mathbf{y} = \left[y(1), \ldots, y(N)\right]^T$, where N is the size of the segment. Let the vectors $\mathbf{x}$ and $\mathbf{w}$ be defined analogously. Let $\mathbf{a}_x = \left(a_{x_0}, \ldots, a_{x_p}\right)$ denote the vector of LP coefficients for the short-time speech segment $\mathbf{x}$ corresponding to $\mathbf{y}$, with $a_{x_0} = 1$ and p the speech LP model order. Similarly, let $\mathbf{a}_w = \left(a_{w_0}, \ldots, a_{w_q}\right)$ denote the LP coefficient vector for the short-time noise segment $\mathbf{w}$ corresponding to $\mathbf{y}$, with $a_{w_0} = 1$ and q being the noise LP model order. Then, the speech and noise PSDs can be written as:

$$P_x(\omega) = \frac{g_x}{|A_x(\omega)|^2} \text{ and } P_w(\omega) = \frac{g_w}{|A_w(\omega)|^2}, \tag{3}$$

where g_x and g_w denote the variance of the prediction error for speech and noise, respectively; $A_x(\omega) = \sum_{k=0}^{p} a_{x_k} e^{-j\omega k}$; and $A_w(\omega) = \sum_{k=0}^{q} a_{w_k} e^{-j\omega k}$. Let

$$\begin{aligned} m_x &= \left[\mathbf{a}_x, g_x\right], \\ m_w &= \left[\mathbf{a}_w, g_w\right]. \end{aligned} \tag{4}$$

m_x is a model describing the speech PSD, and m_w describes the noise PSD. Codebook-driven speech enhancement techniques [25,26] estimate m_x and m_w for each short-time segment: $\mathbf{a}_x$ and $\mathbf{a}_w$ are selected from trained codebooks of vectors of speech and noise LP coefficients, C_x and C_w, respectively, and the gain terms g_x and g_w are computed online, resulting in good performance in nonstationary noise. A maximum likelihood approach is adopted in [25] and a Bayesian minimum mean squared error (MMSE) approach in [26].

The estimates $\hat{m}_x$ and $\hat{m}_w$ are used to construct a Wiener filter to enhance the noisy speech in the frequency domain:

$$H(\omega) = \frac{\hat{P}_x(\omega)}{\hat{P}_x(\omega) + \hat{P}_w(\omega)}, \tag{5}$$

where $\hat{P}_x(\omega)$ and $\hat{P}_w(\omega)$ are estimates of the speech and noise PSDs, respectively, described by $\hat{m}_x$ and $\hat{m}_w$. The Wiener filter is one example of a gain function, and any other gain function can be employed using the obtained speech and noise PSD estimates.

3 Bayesian estimation under varying contexts

In this section, we develop a Bayesian framework to obtain estimates of the speech and noise LP parameters, m_x and m_w, using one or more CD codebooks and a CI speech codebook. The CD codebooks improve estimation accuracy in the event of a context match, and the CI codebook provides robustness in the event of a context mismatch. The Bayesian framework needs to optimally combine the estimates from the various codebooks with no prior knowledge on whether or not the observed context matches the context modeled by the codebooks.

Consider K speech codebooks $\left[C_x^1, \ldots, C_x^K\right]$, which include one or more CD codebooks and a CI codebook, depending on the contexts modeled. We consider a single noise codebook, C_w, corresponding to the encountered noise type. Robustness to different noise types can be provided by extending the notion of context dependency to the noise codebooks as well. To maintain the focus on context dependency in speech, we only consider a single noise codebook.

As m_x is a model for the speech PSD and m_w is a model for the noise PSD, $m = [m_x, m_w]$ is a model for the noisy PSD, given by the sum of the corresponding speech and noise PSDs. We consider m to be a random variable and seek its MMSE estimate, given the noisy observation, the speech codebooks, and the noise codebook. Let $\mathcal{M}_1$ denote the collection of all models of the noisy PSD corresponding to the speech codebook C_x^1 and the noise codebook C_w. The set $\mathcal{M}_1$ consists of quadruplets $\left[\mathbf{a}_x^{1i}, g_x, \mathbf{a}_w^j, g_w\right]$, where $\mathbf{a}_x^{1i}$ is the ith vector from the speech codebook C_x^1, $\mathbf{a}_w^j$ is the jth vector from the noise codebook C_w, and the gain terms g_x and g_w are computed online for each combination of $\mathbf{a}_x^{1i}$ and $\mathbf{a}_w^j$. Thus, $\mathcal{M}_1$ contains $N_x^1 \times N_w$ vectors, where N_x^1 is the number of vectors in C_x^1 and N_w is the number of vectors in C_w. The sets $\mathcal{M}_2, \ldots, \mathcal{M}_K$ are similarly defined, corresponding to the speech codebooks $C_x^2, \ldots, C_x^K$. Let $\mathcal{M}$ be a collection of all the models m contained in all the K speech codebooks and the noise codebook, i.e.,

$$\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2 \cup \ldots \cup \mathcal{M}_K. \tag{6}$$

We consider the following K hypotheses:

- H_k: speech codebook C_k best models the speech context for the current segment, $1 \le k \le K$.

At a given time T, one of the K hypotheses is valid. This corresponds to a state, and we write $S_T = H_k$ to denote that at time T, the most appropriate speech codebook for the observed noisy segment is C_k.

As mentioned in the introductory section, various aspects of context such as speaker, language, etc. can be expected to remain constant over multiple short-time segments, which can be exploited to improve estimation accuracy. The MMSE estimate of m for the Tth short-time segment is thus obtained using not just the current noisy segment $\mathbf{y}_T$ but a sequence that includes the current as well as past noisy segments, $[\mathbf{y}_1, \ldots, \mathbf{y}_T]$, where t is the segment index and $\mathbf{y}_t$, $1 \leq t \leq T$ is a vector containing N noisy speech samples. The MMSE estimate of m can be written as

$$\begin{aligned}\hat{m} &= E\left[m|\mathbf{y}_1, \mathbf{y}_2 \ldots, \mathbf{y}_T\right] \\ &= \sum_{k=1}^{K} p\left(S_T = H_k|\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\right) \\ &\quad \times E\left[m|\mathbf{y}_1, \mathbf{y}_2 \ldots, \mathbf{y}_T, S_T = H_k\right].\end{aligned} \tag{7}$$

The two terms in the last line of (7) lend themselves to an intuitive representation. The second term $E\left[m|\mathbf{y}_1, \mathbf{y}_2 \ldots, \mathbf{y}_T, S_T = H_k\right]$ corresponds to an MMSE estimate of m assuming that the context is best described by H_k. The first term provides a relative importance score to this estimate, based on the likelihood that C_x^k is indeed the most appropriate speech codebook. The weighted summation corresponds to a soft estimation, which allows the coexistence of multiple contexts, e.g., speaker and language, each being modeled by a separate codebook. Next, we derive expressions for both these terms.

First, we consider the term $p\left(S_T = H_k|\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\right)$. Let

$$\alpha_T(k) = p\left(\mathbf{y}_1, \mathbf{y}_2 \ldots, \mathbf{y}_T, S_T = H_k\right),\ k = 1, 2, \ldots, K \tag{8}$$

represent the forward probability as in standard HMM theory [36]. It can be recursively obtained as follows:

Basis step:

$$\alpha_1(k) = p\left(H_k\right) p\left(\mathbf{y}_1|H_k\right),\ k = 1, 2, \ldots, K. \tag{9}$$

The prior probabilities in the absence of any observation can be assumed to be equal in Equation 9. Thus, $p\left(H_k\right) = \frac{1}{K}$, i.e., all hypotheses are equally likely.

Induction step: The state S_T of the current noisy observation $\mathbf{y}_T$ could have been reached from any of the states from the previous frame with a particular transition probability. This can be modeled as

$$\alpha_{t+1}(k) = \left[\sum_{l=1}^{K} \alpha_t(l) a_{lk}\right] p\left(\mathbf{y}_{t+1}|H_k\right), \tag{10}$$

where $1 \leq t \leq T-1$ and $l, k = 1, 2, \ldots, K$, and a_{lk} represent the transition probability of reaching state k from state l. We assume the *a priori* transition probabilities to be known beforehand for a given set of speech codebooks. In this paper, we assume them to be fixed such that a_{lk} takes higher values when $l = k$ than otherwise, to capture the intuition that we typically do not rapidly switch between contexts such as speaker and language. Note that only the *a priori* transition probabilities are assumed to be fixed. The data-dependent part in Equation 10 is captured by the term $p(\mathbf{y}_{t+1}|H_k)$, whose computation is addressed in the following. Using Equation 8,

$$\begin{aligned}p\left(S_T = H_k|\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\right) &= \frac{p\left(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T, S_T = H_k\right)}{p\left(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\right)} \\ &= \frac{\alpha_T(k)}{\sum_{k=1}^{K} \alpha_T(k)}.\end{aligned} \tag{11}$$

Next, we consider the term $E\left[m|\mathbf{y}_1, \mathbf{y}_2 \ldots, \mathbf{y}_T, S_T = H_k\right]$ in Equation 7. In this section, we are interested in exploiting memory to ensure that the codebook that is most relevant to the current context at hand receives a high likelihood, and this is captured by Equation 11. For a given codebook, $E\left[m|\mathbf{y}_1, \mathbf{y}_2 \ldots, \mathbf{y}_T, S_T = H_k\right]$ provides an improved estimate of m by exploiting not only the current noisy observation $\mathbf{y}_T$ but also the past noisy segments. An expression for this term can be derived as in [26], where memory was restricted to the previous frame in view of the signal nonstationarity. Here, to retain the focus on selecting the appropriate context, we assume

$$E\left[m|\mathbf{y}_1, \mathbf{y}_2 \ldots, \mathbf{y}_T, S_T = H_k\right] = E\left[m|\mathbf{y}_T, S_T = H_k\right]. \tag{12}$$

In the following, we ignore the term S_T and write $E[m|\mathbf{y}_T, S_T = H_k]$ as $E[m|\mathbf{y}_T, H_k]$ for brevity. For a given hypothesis H_k, we have

$$\begin{aligned}E\left[m|\mathbf{y}_T, H_k\right] &= \sum_{m \in \mathcal{M}} m\, p\left(m|\mathbf{y}_T, H_k\right) \\ &= \sum_{m \in \mathcal{M}} m\, \frac{p\left(\mathbf{y}_T|m, H_k\right)\ p\left(m|H_k\right)}{p\left(\mathbf{y}_T|H_k\right)}.\end{aligned} \tag{13}$$

Under a Gaussian LP model, m corresponds to an autocorrelation matrix R_y for $\mathbf{y}_T$, which fully characterizes the pdf $p\left(\mathbf{y}_T|m\right)$ as in

$$\begin{aligned}p\left(\mathbf{y}_T|m\right) &= \frac{1}{(2\pi)^{N/2}|R_x + R_w|^{1/2}} \\ &\quad \times \exp\left(-\frac{\mathbf{y}_T^{\dagger}\left(R_x + R_w\right)^{-1}\mathbf{y}_T}{2}\right),\end{aligned} \tag{14}$$

where † represents transpose, $R_y = R_x + R_w$, $R_x = g_x(B_x^{\dagger}B_x)^{-1}$, $R_w = g_w\left(B_w^{\dagger}B_w\right)^{-1}$, B_x is an $N \times N$ lower triangular Toeplitz matrix with $[\mathbf{a}_x, 0, \ldots, 0]^{\dagger}$ as the first column, and B_w is an $N \times N$ lower triangular Toeplitz matrix with $[\mathbf{a}_w, 0, \ldots, 0]^\dagger$ as the first column. Thus, given a model m, $\mathbf{y}_T$ is conditionally independent of H_k, and we have

$$p\left(\mathbf{y}_T|m, H_k\right) = p\left(\mathbf{y}_T|m\right),\ k = 1, 2, \ldots, K. \tag{15}$$

The logarithm of the likelihood $p\left(\mathbf{y}_T|m\right)$ in the Equation 14 can be efficiently computed in the frequency domain following the approach of [26]. The gain terms that maximize the likelihood can be computed as in [26].

Next, we consider the term $p\left(m|H_k\right)$ in Equation 13. Under hypothesis H_k, the speech signal in the observed segment is best described by the codebook C_x^k. We assume all the models resulting from a given codebook are equally likely. This assumption is valid, in general, if the codebook size is large and derived from a phonetically balanced large training set.

Thus, assuming all the models resulting from C_x^k are equally likely, we have

$$\begin{aligned} p\left(m|H_k\right) &= \frac{1}{|\mathcal{M}_k|},\ \forall m \in \mathcal{M}_k \\ &= 0, \text{otherwise}, \end{aligned} \tag{16}$$

where $|\mathcal{M}_k|$ is the cardinality of $\mathcal{M}_k$. From Equations 13 and 16, we have

$$E\left[m|\mathbf{y}_T, H_k\right] = \frac{1}{|\mathcal{M}_k|} \sum_{m \in \mathcal{M}_k} m \frac{p\left(\mathbf{y}_T|m\right)}{p\left(\mathbf{y}_T|H_k\right)}, \tag{17}$$

where

$$p\left(\mathbf{y}_T|H_k\right) = \frac{1}{|\mathcal{M}_k|} \sum_{m \in \mathcal{M}_k} p\left(\mathbf{y}_T|m\right) \tag{18}$$

and $p\left(\mathbf{y}_T|m\right)$ is given by Equation 14. Equation 18 is used in Equations 9 and 10 to obtain the forward probabilities. Finally, the required MMSE estimate $\hat{m}$ is obtained by using Equations 11 and 17 in Equation 7. The speech and noise PSDs corresponding to $\hat{m}$ can be obtained using Equation 3 and the Wiener filter from Equation 5. To ensure stability of the estimated LP parameters, the weighted sum in Equation 7 can be performed in the line spectral frequency domain. Note that the weights are non-negative and add up to unity as is evident from Equation 11. Alternatively, as we are finally interested in the speech and noise PSDs to be used in a Wiener filter, the weighted sum can be performed in the power spectral domain.

We conclude this section with some remarks on the calculation of the forward probabilities α_T which for a codebook captures how well that codebook matches the context of the Tth input segment. As mentioned earlier, the proposed framework can be used to model context in speech as well as noise. When context is modeled by the speech codebooks, it was found to be beneficial to calculate α_T during speech-dominated segments, and during noise-dominated segments when modeling the noise context. The goal in computing α_T is to assess how well a given speech codebook matches the underlying context for a given input segment. If this computation is performed during speech-dominated frames, we obtain accurate values for α_T. However, inaccurate weight values may result when the computation is based on segments that lack sufficient information about the speech, such as silence or low-energy segments dominated by noise. In such situations, it is preferable to use the value of α_T computed in the last speech-dominated segment. This, in other words, assumes that the context of the current segment is the same as that of the past segment. This assumption is valid in general as the context of speech is not expected to rapidly change from one speech burst to another. Thus, updating α_T only during speech-dominated segments does not affect performance. However, estimating α_T only during speech-dominated segments suffers from the disadvantage that there may not be a sufficient number of such segments in highly noisy conditions. Introducing a preliminary noise reduction step, e.g., using the long-term noise estimate from [18], and estimating α_T from the enhanced signal was seen to address this problem. Importantly, the estimation of the speech and noise PSDs and the resulting Wiener filter occurs for each short-time segment, providing good performance under nonstationary noise conditions.

4 Experimental results

Experiments were performed to verify the robustness of the proposed framework under varying contexts. The contexts modeled by a trained CD codebook may or may not match with that of the observed noisy input signal, leading to two scenarios:

- Context match: the best-case scenario for a CD codebook
- Context mismatch: the worst-case scenario for a CD codebook

The robustness of the proposed framework, employing both CD and CI codebooks, was tested under both scenarios. Two different sets of experiments were performed, which differed in terms of number of codebooks employed and the aspects of contexts modeled. The first set consisted of experiments with two speech codebooks, a CI speech codebook and a CD speech codebook, modeling the speaker and acoustic environment as aspects of context. The second set consisted of experiments with three speech codebooks: a CI speech codebook and two CD speech codebooks to study the performance of the proposed framework with an increase in the number of codebooks employed. This set modeled, apart from speaker and acoustic environment, the speech type (normal, whisper, loud, etc.) of the input speech as aspects of context.

In the following, we first describe the experimental setup and, thereafter, the various experiments along with the corresponding results.

4.1 Experimental setup

In all the experiments, the input noisy test utterances were enhanced under different context scenarios, using the CBSE technique [26] applied using the CD codebook alone, the CBSE technique applied using the CI codebook alone, and the proposed Bayesian scheme. We expect that in the context match scenarios, employing the CD codebook alone should lead to the best results. On the other hand, in the context mismatch scenarios, employing the CI codebook alone should lead to results better than those obtained using the CD codebook. The proposed method, however, is expected to provide robust results under varying contexts, i.e., results close to the best results in all scenarios. To serve as a reference for comparisons, we also include results when applying the Wiener filter (5) with a noise estimate obtained from a state-of-the-art noise estimation scheme [37].

The performance of these four processing schemes was compared using two measures: the improvement in segmental SNR (SSNR) referred to as ΔSSNR (in dB) and the improvement in the perceptual evaluation of speech quality (PESQ) [38] measure, referred to as ΔPESQ, averaged over all the enhanced utterances considered under a particular experiment.

The speech codebooks used in the experiments were trained using the Linde-Buzo-Gray (LBG) algorithm [39]. First, the clean speech training utterances, resampled at 8 kHz, were segmented into 50% overlapped Hann windowed frames of size 256 samples each, corresponding to a duration of 32 ms wherein the speech signal can be assumed stationary. Then, LP coefficient vectors of dimension 10, extracted using these frames, were clustered using the LBG algorithm to generate speech codebooks of size 256 each using the Itakura-Saito (IS) distortion [40] as the error criterion.

For training the CI speech codebook, 180 English language utterances of duration 3 to 4 seconds each were used, from 25 male and 25 female speakers from the WSJ speech database [41]. This codebook served as the CI codebook for all the experiments described in this section. The speakers whose utterances were used to train the CI codebook were not used in the test utterances. The different experiments use different CD codebooks and input noisy test data, which are discussed later along with the description of each experiment.

The different CD and CI speech codebooks considered in the experiments are of large size (256) and are derived from a large number of phonetically balanced sentences from the WSJ database. Moreover, the LBG algorithm used to generate the speech codebooks computes cluster centroids in an optimal fashion. All these factors ensure the validity of the assumption about equal probability of models in Equation 16.

Two noise codebooks for two different noise types, traffic and babble, with eight entries each were trained similarly using LP coefficient vectors. For the traffic noise codebook, LP coefficient vectors of order 6 extracted from 2 min of nonstationary traffic noise were used. Since babble noise is speech-like, a higher LP model order of 10 was used while extracting LP coefficient training vectors from approximately 3 min of nonstationary babble noise. The same noise types were also used in the creation of test utterances at 0, 5, and 10 dB SNR for all the experiments. The actual samples were different from those used in training. The active speech level was computed using ITU-T P.56 method B in [42], and noise was scaled and added to obtain a desired SNR.

When processing the noisy files for a particular noise type, the appropriate noise codebook was used. In practice, a classified noise codebook scheme as discussed in [25] can be used. This scheme employs multiple noise codebook, each trained for a particular noise type. A maximum likelihood scheme is used to select the appropriate noise codebook for each short-time frame. This method was shown in [25] to perform as well as the case when the ideal noise codebook was used. We choose to use the ideal noise codebook to retain the focus on the performance of the proposed framework with regard to various aspects of the speech context.

4.2 Experiments with a single CD codebook

In this experiment, we test the proposed framework when two speech codebooks are employed, a CI and a CD codebook. The CD codebook models two aspects of context, 'speaker' and 'acoustic environment'.

4.2.1 CD codebook training

For training the CD codebook, 180 English language utterances from a *single* speaker, of 3 to 4 s duration each, were used from the WSJ speech database. These utterances were convolved with an impulse response recorded at a distance of 50 cm from the microphone, in a reverberant room (T60 = 800 ms). This corresponds, for example, to hands-free mode on a mobile phone. In practice, this codebook is adapted during hands-free usage, making it dependent on both the speaker and acoustic environment.

4.2.2 Test utterances for the experiment

Two sets of ten clean speech utterances each were used to generate the noisy test data. Utterances for the first set were from the same speaker and acoustic environment as the data used to train the CD codebook, corresponding to the context match scenario and thus the best case for the CD codebook. The utterances themselves were different from those used in the training set.

The second set of clean utterances were from a speaker different from the one involved in training the CD codebook. These utterances were not convolved with the recorded impulse response (e.g., corresponding to handset mode in a mobile phone). Thus, both the speaker and acoustic environment were different from those used to train the CD codebook, corresponding to the context mismatch scenario and thus the worst case for the CD codebook.

4.2.3 Enhancement results

The test utterances were enhanced using the four schemes, mentioned in Section 4.1. The transition probabilities a_{lk} were set to 0.99 when $l = k$ and to 0.01 when $l \neq k$, with $l, k = 1, 2$. Tables 1 and 2 provide the results for the best- and worst-case scenarios, respectively, in babble noise.

As can be observed from Table 1, the best results are obtained for the CD codebook, as expected in a context match scenario. There is a significant difference between the results corresponding to the CD and CI codebooks, e.g., 0.19 for ΔPESQ and 1.3 dB for ΔSSNR, at 5 dB input SNR. Moreover, the standard deviation values indicate that the observed differences between the CD and CI results are statistically significant. This illustrates the benefit of employing CD codebooks. On the other hand, Table 2 demonstrates poorer performance when using the CD codebook compared to using the CI codebook, in a context mismatch scenario. The difference between their results is significant for ΔSSNR at all input SNRs, e.g., 1 dB at 0 dB input SNR, and for ΔPESQ at higher SNR, e.g., 0.22 at 10 dB input SNR. These results demonstrate the need for a scheme that appropriately combines the estimates obtained from the CD and CI codebooks, depending on the context at hand.

In Table 1, with increasing input SSNR, there is an increase in ΔPESQ but a decrease in ΔSSNR for all schemes except the reference method. This can be explained by considering the trade-off between speech distortion and noise reduction.

In general, enhancement using a Wiener filter involves applying a gain (also called attenuation) function. When applying this gain function to the noisy speech, both speech and noise components are attenuated. At lower input SNRs, the SSNR measure is dominated by the benefit of noise reduction while ignoring the penalty due to speech distortion. So in these scenarios, applying a greater attenuation than is optimal can increase the output SSNR values as it results in more noise attenuation (it also results in more speech attenuation but that is not captured by the SSNR measure). This situation occurs when using a mismatched codebook, where the clean speech PSD is underestimated, resulting in more severe attenuation of the noisy speech. PESQ is more closer to human perception, and we believe that the effect of speech distortion is better captured by PESQ, resulting in negative delta PESQ values for these scenarios. At higher input SNRs, the SSNR measure also captures the effect of speech distortion. Since ΔPESQ captures well the decrease in speech distortion with increasing input SSNR, there is an increase in ΔPESQ with increasing input SSNR in Table 1. On the other hand, SSNR measure is dominated at lower input SNRs by the benefit of noise reduction ignoring the penalty due to speech distortion. As a result, there is larger ΔSSNR at lower input SNRs than at higher input SNRs.

In contrast to the results obtained when using the CD and CI codebooks alone, the proposed framework achieves robust performance regardless of the observed context. For the best-case scenario (Table 1), its results are close to the CD results. For the worst-case scenario (Table 2), its results are close to the CI results. Thus, the proposed framework achieves results close to the best results for a given scenario, as desired. The reference scheme performs poorly due to the nonstationary nature of the noise. It may be noted that even using a mismatched codebook outperforms the reference scheme, highlighting the benefit of using *a priori* information for speech enhancement in nonstationary noise.

Tables 3 and 4 provide the results for the best- and worst-case scenarios, respectively, for the traffic noise case. Similar observations can be made as from the Tables 1 and 2 regarding the need for both the CI and CD codebooks for better performance and the robust performance of the proposed framework under varying

Table 1 Best-case scenario for a single CD codebook under babble noise

	ΔPESQ			**ΔSSNR (in dB)**		
Input SSNR	**0 dB**	**5 dB**	**10 dB**	**0 dB**	**5 dB**	**10 dB**
CBSE with CD	0.12 ± 0.06	0.18 ± 0.06	0.20 ± 0.07	6.44 ± 0.72	6.01 ± 0.70	4.50 ± 0.88
CBSE with CI	−0.04 ± 0.07	−0.01 ± 0.06	−0.02 ± 0.09	5.59 ± 0.97	4.76 ± 0.92	2.82 ± 1.09
Proposed	0.12 ± 0.06	0.18 ± 0.06	0.20 ± 0.07	6.44 ± 0.72	6.00 ± 0.70	4.49 ± 0.88
Reference	−0.11 ± 0.02	−0.11 ± 0.02	−0.10 ± 0.04	2.08 ± 0.46	2.42 ± 0.47	2.17 ± 0.53

The CD codebook is modeling two aspects of context, speaker and acoustic environment. Both mean and standard deviation values are reported.

Table 2 Worst-case scenario for a single CD codebook under babble noise

	ΔPESQ			ΔSSNR (in dB)		
Input SSNR	**0 dB**	**5 dB**	**10 dB**	**0 dB**	**5 dB**	**10 dB**
CBSE with CD	0.14 ± 0.09	0.12 ± 0.05	0.07 ± 0.05	4.52 ± 0.90	3.72 ± 0.69	1.75 ± 0.68
CBSE with CI	0.17 ± 0.07	0.20 ± 0.06	0.23 ± 0.05	5.51 ± 0.78	5.01 ± 0.74	3.53 ± 0.63
Proposed	0.17 ± 0.07	0.20 ± 0.06	0.21 ± 0.06	5.52 ± 0.79	4.98 ± 0.73	3.47 ± 0.62
Reference	0.09 ± 0.02	0.12 ± 0.03	0.15 ± 0.05	2.40 ± 0.40	2.99 ± 0.40	3.09 ± 0.44

The CD codebook is modeling two aspects of context, speaker and acoustic environment. Both mean and standard deviation values are reported.

contexts. Again, the reference method performs poorly due to the nonstationary nature of noise.

Comparing ΔPESQ values for the best-case scenarios in Tables 1 and 3 for the two noise types shows that there is a sharper drop in values from 5 to 0 dB input SNR in the case of traffic noise results (0.2) compared to babble noise results (0.06). A similar observation can be made for the ΔPESQ values for the worst-case scenarios in Tables 2 and 4 for the two noise types. These observations indicate that the traffic noise case is more difficult to handle than babble noise at 0 dB input SNR. This occurred because the traffic noise considered for the experiments is highly nonstationary compared to the babble noise used for the experiments.

4.2.4 *Comparison of the proposed method with the MMSE-STSA method*

In the above experiments, the reference method chosen for comparison with the proposed method uses the Wiener gain, as described by (5), computed using a state-of-the-art noise estimator [37]. This choice provides an even comparison as the proposed method too employs the Wiener gain function. The two approaches, however, differ in the computation of the speech and noise PSDs for computing the Wiener gain.

Also of interest is a comparison of the proposed method with a popular statistical approach such as the MMSE-STSA method [9], the results of which are provided in Tables 5 and 6 for the Babble noise case. Table 5 corresponds to the context match scenario wherein the context of the CD codebook matches with that of the input noisy speech. Here, the performance of the proposed method is superior, especially for the PESQ values, to that of the MMSE-STSA technique. The advantage with the proposed approach is higher at lower SNR values. For the mismatch scenario, the performance of both the methods is comparable as shown in Table 6. Note that the Wiener filter is just one example of a gain function that can use the speech and noise PSDs estimated using the proposed method. The estimated speech and noise PSDs can also be used to compute the *a priori* and *a posteriori* SNRs for use in the MMSE-STSA gain function. This is however beyond the scope of this paper and is a topic for future work.

4.3 Experiments with multiple CD codebooks

In the previous subsection, we tested the proposed framework under conditions when a single CD codebook was employed along with a CI codebook. Multiple aspects of context were modeled by the single CD codebook. In practice, different contexts will be modeled by different CD codebooks. In this subsection, we experiment with the case of two CD codebooks along with one CI codebook.

4.3.1 *CD codebook training*

The first CD codebook, referred to as CD-1, models a particular speaker and a speech type. The speech type considered is 'whisper' speech. The speech produced in the case of certain speech disorders (dysphonic speech) is similar to whispered speech. CD-1 was trained using around 10 min of whispered speech data from a single speaker from the CHAINS database [43].

The second CD codebook employed, referred to as CD-2, models normal speech in reverberant conditions for the same speaker as modeled by CD-1. CD-2 was trained using training utterances of duration around 10 min,

Table 3 Best-case scenario for a single CD codebook under traffic noise

	ΔPESQ			ΔSSNR (in dB)		
Input SSNR	**0 dB**	**5 dB**	**10 dB**	**0 dB**	**5 dB**	**10 dB**
CBSE with CD	0.00 ± 0.30	0.19 ± 0.06	0.25 ± 0.11	7.84 ± 1.08	6.97 ± 1.11	5.40 ± 1.35
CBSE with CI	−0.21 ± 0.29	−0.05 ± 0.10	0.00 ± 0.14	6.67 ± 1.37	5.57 ± 1.29	3.64 ± 1.49
Proposed	0.01 ± 0.30	0.19 ± 0.06	0.26 ± 0.11	7.83 ± 1.09	6.96 ± 1.11	5.39 ± 1.35
Reference	−0.04 ± 0.31	0.08 ± 0.05	0.08 ± 0.06	2.75 ± 0.49	2.82 ± 0.54	2.21 ± 0.79

The CD codebook is modeling two aspects of context, speaker and acoustic environment. Both mean and standard deviation values are reported.

Table 4 Worst-case scenario for a single CD codebook under traffic noise

	ΔPESQ			ΔSSNR (in dB)		
Input SSNR	**0 dB**	**5 dB**	**10 dB**	**0 dB**	**5 dB**	**10 dB**
CBSE with CD	−0.13 ± 0.10	−0.02 ± 0.11	−0.01 ± 0.07	5.98 ± 1.78	5.28 ± 1.65	3.45 ± 1.63
CBSE with CI	0.09 ± 0.09	0.32 ± 0.09	0.42 ± 0.09	7.35 ± 1.25	7.12 ± 1.25	5.81 ± 1.07
Proposed	0.08 ± 0.09	0.29 ± 0.07	0.39 ± 0.08	7.34 ± 1.24	7.05 ± 1.22	5.75 ± 1.07
Reference	0.21 ± 0.08	0.28 ± 0.10	0.34 ± 0.09	3.21 ± 0.65	3.65 ± 0.65	3.35 ± 0.58

The CD codebook is modeling two aspects of context, speaker and acoustic environment. Both mean and standard deviation values are reported.

convolved with the same impulse response as used in the previous experiments (corresponding to a distance of 50 cm from the microphone, in a reverberant room with T60 = 800 ms).

The two codebooks differ in terms of speaking style, whispered and normal, and also the acoustic environment. The separation in terms of acoustic environment is useful, e.g., to have different CD models for a particular user of the mobile phone to cater to hand-set and hands-free modes of operation. Note that the CI codebook is speaker-independent and corresponds to hand-set mode.

4.3.2 Test utterances for the experiment

Two sets of experiments were performed pertaining to the matching codebook being CD-1 or CD-2. The first set consisted of test utterances generated by adding noise to ten clean 'whispered' speech utterances from the same speaker as in generation of the CD-1 codebook. Similarly, the second set of experiments had test utterances generated using ten clean 'normal' speech utterances from the same speaker as in CD-2, convolved with the same recorded impulse response as used in training CD-2 to constitute the context match scenario for CD-2. In both sets of experiments, the test utterances considered were different from those used in the training of the codebooks. The noisy test utterances were generated as described in the beginning of the section.

4.3.3 Enhancement results

Enhancement using multiple CD codebooks was performed by setting transition probabilities a_{lk} to 0.9 when $l = k$ and to 0.05 when $l \neq k$, with $l, k = 1$ to 3. Tables 7 and 8 present the matching scenario results for CD-1 and CD-2, respectively, for the babble noise case. Similarly, Tables 9 and 10 present the matching scenario results for CD-1 and CD-2, respectively, for traffic noise case. As can be observed from these tables, the best results for all the scenarios occur for the matching CD codebook. The difference between context match and mismatch (between CD-1 and CD-2/CI, and between CD-2 and CD-1/CI) is significant, especially in the ΔPESQ scores. The differences in ΔSSNR values are significant at higher input SNRs. As the number of codebooks employed by the proposed framework increases, there is a possibility of a negative influence from the inappropriate codebooks in the estimation of the model estimate. But from Tables 7, 8, 9, and 10, we observe that for the case of two CD codebooks and one CI codebook, the results for the proposed framework are close to those of the matched codebook at all input SNRs and for both noise types, confirming the robustness of the proposed framework under varying contexts.

5 Conclusions

In this paper, we have introduced the notion of context-dependent (CD) models for speech enhancement methods that use trained models of speech and noise parameters. CD speech models can be trained on one or more aspects of speech context such as speaker, acoustic environment, speaking style, etc., and CD noise models can be trained for specific noise types. Using CD models results in better speech enhancement performance compared to using context-independent (CI) models when the noisy speech shares the same context as the trained codebook. The risk, however, is degraded performance in the event of a context mismatch. Thus, the CD and CI models need to co-exist in a practical implementation. The Bayesian speech enhancement framework proposed

Table 5 Comparison of the proposed method with the MMSE-STSA technique for context match scenario corresponding to Table 1

	ΔPESQ			ΔSSNR (in dB)		
Input SSNR	**0 dB**	**5 dB**	**10 dB**	**0 dB**	**5 dB**	**10 dB**
Proposed	0.12 ± 0.06	0.18 ± 0.06	0.20 ± 0.07	6.44 ± 0.72	6.00 ± 0.70	4.49 ± 0.88
MMSE-STSA	−0.14 ± 0.06	−0.06 ± 0.08	0.04 ± 0.05	5.21 ± 1.28	5.01 ± 0.96	4.21 ± 0.62

Table 6 Comparison of the proposed method with the MMSE-STSA technique for context mismatch scenario corresponding to Table 2

	ΔPESQ			ΔSSNR (in dB)		
Input SSNR	**0 dB**	**5 dB**	**10 dB**	**0 dB**	**5 dB**	**10 dB**
Proposed	0.17 ± 0.07	0.20 ± 0.06	0.21 ± 0.06	5.52 ± 0.79	4.98 ± 0.73	3.47 ± 0.62
MMSE-STSA	0.14 ± 0.04	0.18 ± 0.03	0.23 ± 0.05	5.60 ± 1.14	5.60 ± 0.77	5.03 ± 0.75

Table 7 Results using two CD codebooks and on CI codebook, for context match scenario for CD-1 under babble noise

	ΔPESQ			ΔSSNR (in dB)		
Input SNR	**0 dB**	**5 dB**	**10 dB**	**0 dB**	**5 dB**	**10 dB**
CBSE with CD-1	0.18 ± 0.14	0.18 ± 0.16	0.12 ± 0.14	5.87 ± 1.16	4.88 ± 1.14	2.81 ± 1.27
CBSE with CD-2	0.08 ± 0.18	0.05 ± 0.16	−0.03 ± 0.12	5.69 ± 1.30	4.52 ± 1.17	2.18 ± 1.31
CBSE with CI	0.04 ± 0.17	0.02 ± 0.15	−0.11 ± 0.17	5.41 ± 1.20	4.39 ± 1.16	1.98 ± 1.28
Proposed	0.17 ± 0.13	0.16 ± 0.14	0.07 ± 0.16	5.81 ± 1.14	4.87 ± 1.10	2.58 ± 1.33
Reference	−0.03 ± 0.07	−0.03 ± 0.06	−0.07 ± 0.07	1.76 ± 0.50	1.93 ± 0.43	1.66 ± 0.57

Both mean and standard deviation values are reported.

Table 8 Results using two CD codebooks and one CI codebook, for context match scenario for CD-2 under babble noise

	ΔPESQ			ΔSSNR (in dB)		
Input SNR	**0 dB**	**5 dB**	**10 dB**	**0 dB**	**5 dB**	**10 dB**
CBSE with CD-1	0.11 ± 0.13	0.09 ± 0.09	0.06 ± 0.12	4.15 ± 1.02	3.25 ± 1.18	1.55 ± 1.44
CBSE with CD-2	0.24 ± 0.12	0.21 ± 0.13	0.21 ± 0.12	5.22 ± 1.07	4.64 ± 1.17	3.02 ± 1.32
CBSE with CI	0.18 ± 0.11	0.16 ± 0.10	0.17 ± 0.12	4.77 ± 0.91	4.24 ± 1.02	2.61 ± 1.29
Proposed	0.24 ± 0.12	0.22 ± 0.11	0.21 ± 0.11	5.08 ± 1.12	4.51 ± 1.20	2.93 ± 1.39
Reference	0.08 ± 0.09	0.10 ± 0.07	0.08 ± 0.05	2.59 ± 0.49	3.06 ± 0.51	2.71 ± 0.52

Both mean and standard deviation values are reported.

Table 9 Results using two CD codebooks and one CI codebook, for context match scenario for CD-1 under traffic noise

	ΔPESQ			ΔSSNR (in dB)		
Input SNR	**0 dB**	**5 dB**	**10 dB**	**0 dB**	**5 dB**	**10 dB**
CBSE with CD-1	0.07 ± 0.17	0.24 ± 0.17	0.25 ± 0.18	6.67 ± 1.67	5.70 ± 1.62	3.76 ± 1.50
CBSE with CD-2	−0.16 ± 0.16	−0.03 ± 0.17	−0.03 ± 0.19	6 ± 1.78	4.49 ± 1.82	1.88 ± 2.00
CBSE with CI	−0.1 ± 0.18	0.01 ± 0.16	0.03 ± 0.17	5.85 ± 1.76	4.53 ± 1.87	2.06 ± 1.84
Proposed	0.06 ± 0.16	0.20 ± 0.19	0.22 ± 0.21	6.58 ± 1.68	5.44 ± 1.65	3.19 ± 1.61
Reference	0.05 ± 0.07	0.11 ± 0.09	0.17 ± 0.11	2.54 ± 0.85	2.96 ± 0.91	2.71 ± 1.02

Both mean and standard deviation values are reported.

Table 10 Results using two CD codebooks and one CI codebook, for context match scenario for CD-2 under traffic noise

	ΔPESQ			ΔSSNR (in dB)		
Input SNR	**0 dB**	**5 dB**	**10 dB**	**0 dB**	**5 dB**	**10 dB**
CBSE with CD-1	−0.05 ± 0.13	0.08 ± 0.15	0.13 ± 0.12	6.39 ± 1.40	5.87 ± 0.98	4.4 ± 0.95
CBSE with CD-2	0.01 ± 0.12	0.21 ± 0.15	0.25 ± 0.15	6.69 ± 1.35	6.21 ± 0.91	4.62 ± 0.96
CBSE with CI	−0.07 ± 0.14	0.09 ± 0.16	0.19 ± 0.16	6.48 ± 1.37	5.80 ± 1.02	4.17 ± 1.03
Proposed	0.01 ± 0.12	0.20 ± 0.15	0.27 ± 0.14	6.69 ± 1.36	6.21 ± 0.95	4.70 ± 0.96
Reference	0.07 ± 0.07	0.12 ± 0.10	0.13 ± 0.10	2.76 ± 0.84	3.17 ± 0.83	2.78 ± 0.69

Both mean and standard deviation values are reported.

in this paper obtains estimates of speech and noise parameters based on all available models, requires no prior information on the context at hand, and automatically obtains results close to those obtained when using the appropriate codebook for a given context scenario as seen from experiments with various aspects of speech context.

The improved performance of the proposed method is at the cost of increased computational complexity. As opposed to employing a single CI model, the proposed method involves computations with multiple models. The computations related to each model can, however, occur simultaneously, which allows for a parallel implementation.

The proposed method has been developed using the codebook-based speech enhancement system as an example of a data-driven model-based speech enhancement system. Other model-based schemes, such as those using HMMs, GMMs, and NMF, can benefit in a similar manner, and the extension is a topic for future work. The theory developed in this paper is directly applicable to context-dependent noise codebooks and can be used for robust noise estimation under varying noise conditions.

In this paper, context-dependent models are assumed to be available. In practice, they need to be trained online. For several aspects of context, a separate enrollment stage may not be meaningful and the models need to be progressively adapted during usage when the SNR is high. Distinguishing between different aspects of context and training separate models for them online is another topic for future work.

The codebooks considered in this paper consist of vectors of tenth-order LP coefficients, which model the smoothed spectral envelope. It will be worthwhile to investigate the suitability of other spectral representations such as higher resolution PSDs, mel-frequency cepstral coefficients, etc., to capture context-dependent information. Different features may be employed depending on which aspects of context are to be modeled and depending on the application, e.g., whether the enhancement is for speech communication, speaker identification, or for speech recognition.

Competing interests

The authors declare that they have no competing interests.

Authors' information

This work was performed when SS was with Philips Research Laboratories, Eindhoven, The Netherlands.

Acknowledgements

The authors would like to thank Prof. G. V. Prabhakara Rao, Head, Department of Information Technology, Rajiv Gandhi Memorial College of Engineering and Technology, Nandyal, Andhra Pradesh, India, for valuable discussions on this topic.

Author details

[1]Department of Mathematics and Computer Science, Sri Sathya Sai Institute of Higher Learning, Prasanthi Nilayam, Anantapur, Andhra Pradesh 515134, India.
[2]Microsoft Corporation, Redmond, WA 98052, USA.

Received: 31 October 2013 Accepted: 18 August 2014
Published online: 12 September 2014

References

1. B Schuller, M Wöllmer, T Moosmayr, G Rigoll, Recognition of noisy speech: a comparative survey of robust model architecture and feature enhancement. EURASIP J. Audio Speech Music Process. **2009**, 1–17 (2009)
2. V Hamacher, J Chalupper, J Eggers, E Fischer, U Kornagel, H Puder, U Rass, Signal processing in high-end hearing aids: state of the art, challenges, and future trends. EURASIP J. Appl. Signal Process. **2005**(18), 2915–2929 (2005)
3. SF Boll, Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. **27**(2), 113–120 (1979)
4. M Berouti, M Schwartz, J Makhoul, Enhancement of speech corrupted by acoustic noise, in *Proceedings of the IEEE Int. Conf. Acoust. Speech Signal Processing (ICASSP)* (Washington D. C., 2–4 April 1979), pp. 208–211
5. S Kamath, P Loizou, A multi-band spectral subtraction method for enhancing speech corrupted by colored noise, in *Proceedings of the IEEE Int. Conf. Acoust. Speech Signal Processing (ICASSP)* (Orlando, 13–17 May 2002), pp. IV-4164
6. Y Lu, PC Loizou, A geometric approach to spectral subtraction. Speech Commun. **50**(6), 453–466 (2008)
7. K Paliwal, B Schwerin, K Wojcicki, Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator. Speech Commun. **54**(2), 282–305 (2012)
8. RJ McAulay, ML Malpass, Speech enhancement using a soft-decision noise suppression filter. IEEE Trans. Acoust. Speech Signal Process. **28**(2), 137–145 (1980)
9. Y Ephraim, D Malah, Speech enhancement using a minimum mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **32**(6), 1109–1121 (1984)
10. E Plourde, B Champagne, Multidimensional STSA estimators for speech enhancement with correlated spectral components. IEEE Trans. Sig. Proc. **59**(7), 3013–3024 (2011)
11. BJ Borgstrom, A Alwan, A unified framework for designing optimal STSA estimators assuming maximum likelihood phase equivalence of speech and noise. IEEE Trans. Audio Speech Language Process. **19**(8), 2579–2590 (2011)
12. Y Andrianakis, PR White, Speech enhancement algorithm based on a Chi MRF of the speech STFT amplitudes. IEEE Trans. Acoust. Speech Signal Process. **17**(8), 1508–1517 (2009)
13. M McCallum, B Guillemin, Stochastic-deterministic MMSE STFT speech enhancement with general a priori information. IEEE Trans. Audio Speech Language Process. **21**(7), 1445–1457 (2013)
14. Y Ephraim, HL Van Trees, A signal subspace approach for speech enhancement. IEEE Trans. Acoust. Speech Signal Process. **3**(4), 251–266 (1995)
15. P Loizou, *Speech Enhancement: Theory and Practice*. (CRC Press, Boca Raton, 2007)
16. K Srinivasan, A Gersho, Voice activity detection for cellular networks, in *Proceedings of the IEEE Speech Coding Workshop* (Sainte-Adèle, 13–15 October 1993), pp. 85–86
17. J Gorriz, J Ramirez, E Lan, C Puntonet, Jointly Gaussian pdf-based likelihood ratio test for voice activity detection. IEEE Trans. Audio Speech Language Process. **16**(8), 1565–1578 (2009)
18. R Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. **9**(4), 504–512 (2001)
19. I Cohen, Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Trans. Acoust. Speech Signal Process. **11**(5), 466–475 (2003)
20. JS Erkelens, R Heusdens, Tracking of nonstationary noise based on data-driven recursive noise power estimation. IEEE Trans. Audio, Speech, Language Process. **16**(6), 1112–1123 (2008)

21. M Souden, M Delcroix, K Kinoshita, T Yoshioka, T Nakatani, Noise power spectral density tracking: a maximum likelihood perspective. IEEE Sig. Process. lett. **19**(8), 495–498 (2012)
22. R Hendriks, R Heusdens, J Jensen, MMSE based noise PSD tracking with low complexity, in *Proc. of IEEE International Conf. on Acoustics Speech and Signal Processing (ICASSP), 2010* (Dallas, 14–19 March 2010), pp. 4266–4269
23. T Gerkmann, R Hendriks, Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. IEEE Trans. Audio, Speech, Language Process. **20**(4), 1383–1393 (2012)
24. TV Sreenivas, P Kirnapure, Codebook constrained Wiener filtering for speech enhancement. IEEE Trans. Acoust. Speech Signal Process. **4**(5), 383–389 (1996)
25. S Srinivasan, J Samuelsson, WB Kleijn, Codebook driven short-term predictor parameter estimation for speech enhancement. IEEE Trans. Audio Speech Language Process. **14**(1), 163–176 (2006)
26. S Srinivasan, J Samuelsson, WB Kleijn, Codebook-based Bayesian speech enhancement for nonstationary environments. IEEE Trans. Audio Speech Language Process. **15**(2), 441–452 (2007)
27. X Xiao, RM Nickel, Speech enhancement with inventory style speech resynthesis. IEEE Trans. Audio, Speech Language Process. **18**(6), 1243–1257 (2010)
28. T Rosenkranz, H Puder, Improving robustness of codebook-based noise estimation approaches with delta codebooks. IEEE Trans. Audio Speech Language Process. **20**(4), 1177–1188 (2012)
29. H Sameti, H Sheikhzadeh, L Deng, HMM-based strategies for enhancement of speech signals embedded in nonstationary noise. IEEE Trans. Acoust. Speech Signal Process. **6**(5), 445–455 (1998)
30. DY Zhao, WB Kleijn, HMM-based gain-modeling for enhancement of speech in noise. IEEE Trans. Audio Speech Language Process. **15**(3), 882–892 (2007)
31. H Veisi, H Sameti, Speech enhancement using hidden Markov models in Mel-frequency domain. Speech Commun. **55**(2), 205–220 (2013)
32. J Hao, T-W Lee, TJ Sejnowski, Speech enhancement using Gaussian scale mixture models. IEEE Trans. Audio Speech Language Process. **18**(6), 1127–1136 (2010)
33. N Mohammadiha, P Smaragdis, A Leijon, Supervised and unsupervised speech enhancement using nonnegative matrix factorization. IEEE Trans. Audio Speech Language Process. **21**(10), 2140–2151 (2013)
34. C Sigg, T Dikk, J Buhmann, Speech enhancement using generative dictionary learning. IEEE Trans. Audio, Speech, Language Process. **20**(6), 1698–1712 (2012)
35. DHR Naidu, S Srinivasan, A Bayesian framework for robust speech enhancment under varying contexts, in *Proceedings of the IEEE Int. Conf. Acoust. Speech Signal Processing (ICASSP)* (Kyoto, 25–30 March 2012), pp. 4557–4560
36. LR Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE. **77**(2), 257–286 (1989)
37. S Rangachari, P Loizou, A noise estimation algorithm for highly nonstationary environments. Speech Commun. **28**, 220–231 (2006)
38. A Rix, J Beerends, M Hollier, A Hekstra, Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs, in *Proceedings of the IEEE Int. Conf. Acoust. Speech Signal Processing (ICASSP)* (Salt Lake City, 7–11 May 2001), pp. 749–752
39. Y Linde, A Buzo, RM Gray, An algorithm for vector quantizer design. IEEE Trans. Commun. **28**(1), 84–95 (1980)
40. R Gray, A Buzo, A Gray, Y Matsuyama, Distortion measures for speech processing. IEEE Trans. Acoust. Speech Signal Process. **28**(4), 367–376 (1980)
41. CSR-II (WSJ1) Complete LDC94S13A. DVD. Philadelphia: Linguistic Data Consortium (1994)
42. ITU-T Rec. P.56, Objective measurement of active speech level. International Telecommunication Union, CH-Geneva (1993)
43. F Cummins, M Grimaldi, T Leonard, J Simko, The CHAINS corpus: characterizing individual speakers, in *Proceedings of the International Conference on Speech and Computer (SPECOM)* (St Petersburg, 2006), pp. 431–435

doi:10.1186/s13636-014-0035-4
Cite this article as: Naidu and Srinivasan: **Robust Bayesian estimation for context-based speech enhancement.** *EURASIP Journal on Audio, Speech, and Music Processing* 2014 **2014**:35.