A BAYESIAN FRAMEWORK FOR ROBUST SPEECH ENHANCEMENT UNDER VARYING CONTEXTS

D. Hanumantha Rao Naidu

Dept. Math and Comp. Science
Sri Sathya Sai Institute of Higher Learning
Prasanthi Nilayam, India
dhanumantharao@sssihl.edu.in

Sriram Srinivasan

Digital Signal Processing Group
Philips Research
Eindhoven, The Netherlands
sriram.srinivasan@philips.com

ABSTRACT

Single-microphone speech enhancement algorithms that employ trained codebooks of parametric representations of speech spectra have been shown to be successful in the suppression of non-stationary noise, e.g., in mobile phones. In this paper, we introduce the concept of a context-dependent codebook, and look at two aspects of context: dependency on the particular speaker using the mobile device, and on the acoustic condition during usage (e.g., hands-free mode in a reverberant room). Such context-dependent codebooks may be trained on-line. A new scheme is proposed to appropriately combine the estimates resulting from the context-dependent and context-independent codebooks under a Bayesian framework. Experimental results establish that the proposed approach performs better than the context-independent codebook in the case of a context match and better than the context-dependent codebook in the case of a context mismatch.

Index Terms— Speech enhancement, noise reduction, context-dependent, codebook, linear prediction

1. INTRODUCTION

Single channel speech enhancement is a challenging problem with several important applications [1]. With the widespread use of mobile communication, the problem of enhancing noisy speech has gained substantial importance. As background noise is common in typical conversation environments such as in traffic, restaurants etc., speech enhancement has become an important component in mobile phones. Enhancement of speech is also relevant in the use of hearing aids [2], and has a significant role as a pre-processor for providing a clean input to speech recognition systems whose performance is affected by the noise level in the input [3] .

There exist several single-channel speech enhancement algorithms in the literature [1]. Most of them work well in the case of stationary noise. However, their performance severely degrades when dealing with non-stationary noise. Codebook-based speech enhancement techniques [4, 5] have been proven to perform well in the presence of non-stationary noise. These techniques rely on trained codebooks of speech and noise LP vectors. The early codebook-based methods employed speech codebooks trained using several speakers, and are speaker independent (SI) in nature. Speaker dependent (SD) codebooks, which are codebooks generated using speech training data from a single speaker, were introduced in [6]. Such SD codebooks are relevant in the context of mobile communications and/or certain speech recognition systems where the device is used by a single user for most of the time, and may be trained on-line. It has been shown that speaker dependency in the SD codebooks also translates into a noticeable improvement in speech enhancement when compared to using SI codebooks in the codebook based Bayesian speech enhancement framework [6].

The contribution of this paper is two-fold. Firstly, we generalize the notion of speaker dependency to that of context dependency, and consider two aspects of the context: the speaker using the device, and the acoustic conditions in the room. Secondly, we introduce a new Bayesian estimation approach to provide a solution that can exploit context dependency while remaining robust to a context mismatch.

When a speech codebook is adapted on-line, e.g., during periods when there is little or no noise, the codebook is not only adapted to the particular speaker's voice but to the entire context, which may include the prevailing acoustic conditions, the microphone characteristics, etc. The acoustic conditions will vary depending on whether the phone is in hands-free or hand-set mode, and when in hands-free mode, the conditions are determined by the amount of reverberation in the room. In this paper, we restrict the context to include the speaker's voice and the prevailing acoustic conditions as these may be expected to have the largest impact on performance. In the absence of context-specific information, the default codebook is speaker-independent and trained for hand-set operation, where the speaker is close to the microphone.

It is intuitive to expect that a context-dependent (CD) codebook will provide significantly better performance than a context-independent (CI) codebook. However, it is important that a method that exploits context dependency is robust to a context mismatch, e.g., when used by a different person or in a different acoustic condition. As codebook adaptation can take place only when the signal-to-noise ratio (SNR) is sufficiently high, it is not possible to adapt the speech codebook to the new context if the mismatch occurs in a noisy environment. In a practical system, the CD and CI codebooks will need to co-exist.

In this paper, we propose a new Bayesian approach that provides a framework to combine the estimates resulting from the CD and the CI codebook in a manner that exploits the benefits of both codebooks, and is robust to a mismatch between the context on which the CD codebook is trained and the context that is encountered during testing. Ideally, such a method should result in an estimate that is better than the estimate obtained using a CI codebook in the event of a context match, and better than the estimate obtained using a CD codebook in the event of a context mismatch.

The remainder of the paper is organized as follows. Section 2 describes the signal model. The proposed Bayesian approach is developed in Section 3. Section 4 presents experimental results followed by conclusions in Section 5.

2. SIGNAL MODEL

We consider an additive noise model, where the noisy signal $y(n)$ can be written as

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

where n is the time index, $x(n)$ is the clean speech signal and $w(n)$ is the noise signal. Let $\mathbf{y} = [y(1), \ldots, y(N)]^T$ denote a vector of noisy observations of length N, corresponding to a short-time segment.

Prior information about the speech and noise signals is captured in the form of trained codebooks of speech and noise autoregressive (AR) parameters. For a given short time segment $\mathbf{y}$, let $\theta_x = (a_{x_0}, \ldots, a_{x_p})$ denote the vector of speech AR parameters, where $a_{x_0} = 1$ and p is the speech AR model order, and let g_x denote the variance of the speech excitation signal. Similarly, let $\theta_w = (a_{w_0}, \ldots, a_{w_q})$ denote the vector of noise AR parameters, where $a_{w_0} = 1$ and q is the noise AR model order, and let g_w denote the variance of the noise excitation signal.

Let $m = [\theta_x, \theta_w, g_x, g_w]$. The goal of codebook-based speech enhancement algorithms is to obtain an estimate of m using the speech and noise codebooks, and given the noisy observation. A maximum-likelihood estimate is obtained in [4] and a Bayesian minimum mean-squared error (MMSE) estimate in [5]. Once such an estimate is obtained, it can be used to construct an estimate of the speech and noise power spectral densities (PSDs) as follows:

$$\hat{P}_x(\omega) = \frac{g_x}{|A_x(\omega)|^2} \text{ and } \hat{P}_w(\omega) = \frac{g_w}{|A_w(\omega)|^2}, \tag{2}$$

where $A_x(\omega) = \sum_{k=0}^{p} a_{x_k} e^{-j\omega k}$ and $A_w(\omega) = \sum_{k=0}^{q} a_{w_k} e^{-j\omega k}$. These PSDs can now be used to construct a Wiener filter to enhance the noisy speech in the frequency domain:

$$H(\omega) = \frac{\hat{P}_x(\omega)}{\hat{P}_x(\omega) + \hat{P}_w(\omega)}. \tag{3}$$

In addition to exploiting prior information about speech and noise, the benefit of the codebook-based approaches lies in the frame-by-frame estimation of the optimal gain levels g_x and g_w, which results in good performance under practical non-stationary noise conditions.

As explained in Section 1, estimation accuracy can be improved by considering speech codebooks that are context-dependent. In the following section, a new Bayesian estimation approach is developed that obtains the MMSE estimate of m given the noisy observation, a CD speech codebook, a CI speech codebook and a noise codebook. By casting the problem in a Bayesian framework, estimates from the context dependent and independent codebooks are automatically combined in an optimal (in the MMSE sense) manner given the noisy observation.

3. BAYESIAN AR PARAMETER ESTIMATION

As explained in Section 2, the random variable m represents a model describing the speech and noise PSDs that constitute the observed noisy PSD. We seek an expression for $\hat{m} = E[m|\mathbf{y}]$. Consider the following two hypotheses:

- H_0 : CD codebook is the appropriate codebook
- H_1 : CI codebook is the appropriate codebook

The MMSE estimate of m can be written as

$$\hat{m} = E[m|\mathbf{y}] = \sum_{k=0}^{1} p(H_k|\mathbf{y}) E[m|\mathbf{y}, H_k] \tag{4}$$

Let $\mathcal{M}$ be the collection of all models. In this paper, $\mathcal{M} = \mathcal{M}_{\text{CD}} \cup \mathcal{M}_{\text{CI}}$, where $\mathcal{M}_{\text{CD}}$ is the collection of all context dependent models and $\mathcal{M}_{\text{CI}}$ is the collection of all context independent models. In the following, we assume that only the speech codebook is adapted to the context. The extension to the case where the noise codebook is also context dependent is straightforward. The set $\mathcal{M}_{\text{CD}}$ consists of all quadruplets $[\theta_x^i, \theta_w^j, g_x, g_w]$, where θ_x^i is the i^{th} entry from the CD speech codebook, and θ_w^j is the j^{th} entry from the noise codebook. The gain terms, as mentioned earlier, are computed online for each combination of θ_x^j and θ_w^j. Thus, $\mathcal{M}_{\text{CD}}$ has $N_{\text{CD}} \times N_w$ models, where N_{CD} is the number of entries in the CD speech codebook and N_w is the number of entries in the noise codebook. The set $\mathcal{M}_{\text{CI}}$ is constructed analogously and has $N_{\text{CI}} \times N_w$ models, where N_{CI} is the number of entries in the CI speech codebook. We have for $k = 0, 1$

$$\begin{aligned} E[m|\mathbf{y}, H_k] &= \sum_{m \in \mathcal{M}} m\, p(m|\mathbf{y}, H_k) \\ &= \sum_{m \in \mathcal{M}} m\, \frac{p(\mathbf{y}|m, H_k)\, p(m|H_k)}{p(\mathbf{y}|H_k)} \end{aligned} \tag{5}$$

Given a model m, $\mathbf{y}$ is conditionally independent of H_k. Thus,

$$p(\mathbf{y}|m, H_k) = p(\mathbf{y}|m),\ k = 0, 1. \tag{6}$$

Under a Gaussian AR model, the likelihood $p(\mathbf{y}|m)$ is given by

$$p(\mathbf{y}|m) = \frac{1}{(2\pi)^{N/2}|R_x + R_w|^{1/2}} \exp\left(-\frac{\mathbf{y}^T (R_x + R_w)^{-1} \mathbf{y}}{2}\right), \tag{7}$$

where $R_x = g_x (B_x^T B_x)^{-1}$, $R_w = g_w (B_w^T B_w)^{-1}$, B_x is the $N \times N$ lower triangular Toeplitz matrix with $[\theta_x, 0, \ldots, 0]^T$ as the first column and B_w is the $N \times N$ lower triangular Toeplitz matrix with $[\theta_w, 0, \ldots, 0]^T$ as the first column. The logarithm of the likelihood $p(\mathbf{y}|m)$ can be efficiently computed in the frequency domain following the approach of [5]. The gain terms that maximize the likelihood can be computed as in [5].

Next, we consider the term $p(m|H_k)$ in equation (5). Under H_0, the speech signal in the observed segment is best described by the CD codebook, and therefore we have

$$\begin{aligned} p(m|H_0) &= \frac{1}{|\mathcal{M}_{\text{CD}}|},\ \forall m \in \mathcal{M}_{\text{CD}} \\ &= 0,\ \text{otherwise.} \end{aligned} \tag{8}$$

where $|\mathcal{M}_{\text{CD}}|$ is the cardinality of $\mathcal{M}_{\text{CD}}$, and we assumed that all context dependent models are equally likely. Similarly,

$$\begin{aligned} p(m|H_1) &= \frac{1}{|\mathcal{M}_{\text{CI}}|},\ \forall m \in \mathcal{M}_{\text{CI}} \\ &= 0,\ \text{otherwise.} \end{aligned} \tag{9}$$

where $|\mathcal{M}_{\text{CI}}|$ is the cardinality of $\mathcal{M}_{\text{CI}}$. From (5) and (8), we have

$$E[m|\mathbf{y}, H_0] = \frac{1}{|\mathcal{M}_{\text{CD}}|} \sum_{m \in \mathcal{M}_{\text{CD}}} m \frac{p(\mathbf{y}|m)}{p(\mathbf{y}|H_0)} \tag{10}$$

where

$$p(\mathbf{y}|H_0) = \frac{1}{|\mathcal{M}_{\text{CD}}|} \sum_{m \in \mathcal{M}_{\text{CD}}} p(\mathbf{y}|m) \tag{11}$$

Similarly, from (5) and (9), we have

$$E[m|\mathbf{y}, H_1] = \frac{1}{|\mathcal{M}_{\text{CI}}|} \sum_{m \in \mathcal{M}_{\text{CI}}} m \frac{p(\mathbf{y}|m)}{p(\mathbf{y}|H_1)} \tag{12}$$

where

$$p(\mathbf{y}|H_1) = \frac{1}{|\mathcal{M}_{\mathrm{CI}}|} \sum_{m \in \mathcal{M}_{\mathrm{CI}}} p(\mathbf{y}|m) \tag{13}$$

To obtain $\hat{m}$ using (4), now only $p(H_k|\mathbf{y})$ needs to be determined, which can be obtained as

$$p(H_k|\mathbf{y}) = \frac{p(\mathbf{y}|H_k)p(H_k)}{p(\mathbf{y})},\ k = 0, 1, \tag{14}$$

where

$$p(\mathbf{y}) = \sum_{k=0}^{1} p(\mathbf{y}|H_k) \tag{15}$$

and $p(\mathbf{y}|H_0)$ and $p(\mathbf{y}|H_1)$ are given by the equations (11) and (13), respectively. The prior probabilities in the absence of any observation are assumed to be equal so that $p(H_0) = p(H_1) = 0.5$. The MMSE estimate $\hat{m}$ is obtained using (10), (12), and (14) in (4). The speech and noise PSDs corresponding to $\hat{m}$ can be obtained using (2), and the Wiener filter from (3).

4. EXPERIMENTAL RESULTS

In this section, we present the details and results of the experiments performed with the following two objectives. Firstly, to investigate the benefits of context dependency and secondly, to validate the robustness of the proposed Bayesian approach when there is a mismatch between the training and testing context. As mentioned in Section 1, in this paper, context refers to the speaker and acoustic conditions. Thus, the CD codebook is trained for a particular speaker's data and in a hands-free recording set-up. The CI codebook is trained with data from several speakers and in a hand-set recording set-up. In practice, the CI codebook would be shipped with the mobile device and the CD codebook would result from adaptation during usage. For the experimental analysis in this paper, it is assumed that the CD codebook has been fully adapted.

4.1. Codebook training

CI and CD codebooks of speech AR coefficients were generated using training data from Wall Street Journal (WSJ) speech database [7]. 180 distinct training utterances of duration around 5 sec each were used in generation of each of the codebooks. In both the cases, the speech content was the same while the number of speakers used for training varied. To generate the CI codebook, utterances from 50 different speakers, 25 male and 25 female were used. The training utterances for the CD codebook were from a single male speaker, and were convolved with a recorded impulse response at a distance of 50 cm from the microphone in a reverberant room (T60 = 800 ms). Using the training utterances, around 55000 LP coefficients were extracted from Hann-windowed segments of size 256 samples, with a 50 percent overlap at a sampling frequency of 8 kHz. The codebooks were trained using LBG algorithm [8] with the root mean squared log-spectral distortion (LSD) as the error criterion. The codebook size was fixed at 256 entries. Larger codebooks did not result in a significant increase in performance.

The noise codebook was trained in a similar fashion. To deal with different noise types when using trained noise codebooks, a classified noise codebook scheme can be employed as in [4]. In this paper, we assume that the correct noise codebook is available, to retain the focus on the study of the benefit and robustness of the proposed Bayesian scheme to exploit context dependency.

4.2. Test scenarios

The test data consisted of noisy files generated by adding babble noise at an SNR of 5 dB to ten clean utterances from the WSJ database. The content of the test utterances was different from that of the training utterances. Depending on the test scenario, the utterance were selected from either the same speaker (referred to in the following as speaker A) whose data was used to train the CD codebook or from a different speaker (referred to in the following as speaker B). Furthermore, again depending on the test scenario, the test data was either convolved with the impulse response corresponding to the hands-free (HF) recording conditions considered in the training, or to the hand-set (HS) condition. Four test scenarios were considered as tabulated below.

Scenario	Speaker type	Recording condition
1	A	HF
2	B	HS
3	B	HF
4	A	HS

Table 1. Composition of the test files in the different test scenarios. The CD codebook is trained on data from speaker A and in HF mode. The CI codebook is trained on data from speaker B and in HS mode.

The input noisy files under each test scenario were enhanced using the CD codebook alone, the CI codebook alone, and using the proposed Bayesian approach that optimally combines the estimates resulting from the CD and CI codebooks. Clearly, for certain scenarios in Table 1, the CD codebook may be expected to provide the best performance, and for certain others the CI codebook will perform well. We expect the proposed approach to be robust in all scenarios; specifically, for each scenario, its performance should be better than the worst of the two (CD or CI), and close to the performance of the best of the two.

The performance of these three processing schemes was quantified using the improvement in segmental SNR (SSNR) referred to as ΔSSNR (in dB) and the improvement in the PESQ [9] measure, referred to as ΔPESQ, averaged over the ten files. While PESQ was not originally developed for evaluating the performance of speech enhancement algorithms, it has been shown to have a good correlation to subjective quality, and including these results also serves to validate PESQ as a measure in this context.

4.2.1. Test files from Speaker A in HF mode

This is the best-case scenario for the CD codebook as it is adapted to speaker A in HF mode, and the results are shown in Table 2.

	CD	CI	Proposed
ΔSSNR	5.9	4.1	5.2
ΔPESQ	0.17	-0.03	0.11

Table 2. Results for the best-case scenario. Both the speaker and acoustic conditions are identical to those for the CD codebook.

Clearly, the CD codebook performs better than the CI codebook for both the measures, with a difference of 1.8 dB in ΔSSNR and 0.2 in the case of ΔPESQ, illustrating the benefit of context dependency. Note that ΔSSNR and ΔPESQ are not absolute values but the improvement compared to the noisy input. The proposed Bayesian

approach adapts to the context and its performance is closer to that of the CD codebook, as expected. The difference in performance between the proposed approach and CD shows the penalty to be paid for the ensuring robustness to a context mismatch. The benefit of this robustness will become apparent in the next test scenario. In any case, the proposed approach performs significantly better than CI, by around 1.1 dB in ΔSSNR and 0.1 in ΔPESQ.

4.2.2. Test files from Speaker B in HS mode

This is the worst-case scenario for the CD codebook, and Table 3 summarizes the results. The results for CD are poor compared to CI,

	CD	CI	Proposed
ΔSSNR	3.0	4.8	4.2
ΔPESQ	0.12	0.22	0.22

Table 3. Results for the worst-case scenario. Both the speaker and acoustic conditions are different from those for the CD codebook.

as expected. Once again, the proposed approach adapts according to the context and its performance is close to that of the CI codebook. The benefit of the proposed codebook over CD in the event of a context mismatch is 1.2 dB in ΔSSNR and 0.1 in ΔPESQ. The current and previous test scenarios summarize the robust performance of the proposed approach in the event of a context match as well as a context mismatch.

4.2.3. Test files from Speaker B in HF mode

This scenario corresponds to a partial context match, where the test files are from a different speaker than the one on which the CD codebook is adapted, but the recordings conditions match. Table 4 shows

	CD	CI	Proposed
ΔSSNR	3.1	4.3	4.0
ΔPESQ	0.02	0.1	0.1

Table 4. Results for a partial context match. The acoustic conditions are identical to those for the CD codebook, but the speaker is different.

that the CI codebook performs better than the CD codebook in this case, highlighting the dominance of the effect of accurate speaker modeling over the dependency on acoustic conditions. Again, the proposed approach automatically adapts to the context and its performance is close to that of the CI codebook.

4.2.4. Test files from Speaker A in HS mode

The final scenario also corresponds to a partial context match, where the test files are from the same speaker as the one on which the CD codebook is adapted, but the recordings conditions do not match. Unlike in the previous experiment, the performance of the CI codebook is poorer than that of the CD codebook. This re-emphasizes the observation that among the two aspects of context considered here, speaker dependency plays a more dominant role in codebook-based speech enhancement. Once again we observe the adaptability of the proposed Bayesian approach which leans towards the better performance of the CD codebook.

	CD	CI	Proposed
ΔSSNR	6.0	5.1	5.6
ΔPESQ	0.23	0.14	0.21

Table 5. Results for a partial context match. The speaker is the same as that for the CD codebook, but the acoustic conditions are different.

4.2.5. Summary of test results

The four scenarios considered here demonstrate the effect of context on the performance of the codebook-based speech enhancement algorithm. Depending on the particular scenario, either the CD or the CI codebook performs better, with a significant difference between the two, highlighting the need for a robust method that performs well in all scenarios. The proposed Bayesian approach provides this desired robustness, and its performance in all four scenarios was seen to be close to that of the best performing codebook. While the results reported here are for an input SNR of 5 dB, these trends were also observed at other input SNRs.

5. CONCLUSIONS

The concept of context-dependent (CD) codebooks for codebook-based speech enhancement was introduced. Two aspects of context dependency were considered for investigating the benefits of the CD codebooks over context-independent (CI) codebooks: dependency on the speaker characteristics, and on the acoustic condition during usage. A Bayesian scheme was proposed for providing a framework to optimally combine the estimates from the CD and CI codebooks. The scheme exploits the benefits of context dependency, and is also robust to a mismatch between the context on which the CD codebook is trained and the context that is encountered during testing. Experimental results under different test scenarios confirm the benefits and robustness of the proposed approach.

6. REFERENCES

[1] P. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2007.

[2] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass, "Signal processing in high-end hearing aids: State of the art, challenges, and future trends," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 18, pp. 2915–2929, 2005.

[3] B. Schuller, M. Wöllmer, T. Moosmayr, and G. Rigoll, "Recognition of noisy speech: a comparative survey of robust model architecture and feature enhancement," *EURASIP J. Audio Speech Music Process.*, vol. 2009, pp. 1–17, 2009.

[4] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 14(1), pp. 163–176, 2006.

[5] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Trans. Speech Audio Process.*, vol. 15(2), pp. 441–452, 2007.

[6] D. H. R. Naidu, G. V. P. Rao, and S. Srinivasan, "Speech enhancement using speaker dependent codebooks," *17th International Conference on Digital Signal Processing (DSP 2011)*, vol. 2011, pp. 1–6, 2011.

[7] "CSR-II (WSJ1) Complete," *Linguistic Data Consortium*, Philadelphia, 1994.

[8] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. 28(1), pp. 84–95, 1980.

[9] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 2, pp. 749–752, 2001.