Speech enhancement using a generic noise codebook

Sriram Srinivasan*, Member, IEEE, and D. Hanumantha Rao Naidu, Student Member, IEEE

Abstract

Although single-microphone noise reduction methods perform well in stationary noise environments, their performance in non-stationary conditions remains unsatisfactory. Use of prior knowledge about speech and noise power spectral densities in the form of trained codebooks has been previously shown to address this limitation. While it is possible to use trained speech codebooks in a practical system, the variety of noise types encountered in practice makes the use of trained noise codebooks less practical. This letter presents a new approach that uses a generic noise codebook for speech enhancement that can be generated on-the-fly and provides good performance.

Index Terms

Speech enhancement, codebook, noise reduction.

EDICS Category: SPE-ANAL Speech coding, synthesis and analysis

S. Srinivasan is with Philips Research, High Tech Campus 36, WO-2, 5656AE Eindhoven, The Netherlands. Tel: +31 40 2746157 email: sriram.srinivasan@philips.com

D. H. R. Naidu is with the Dept. of Mathematics and Computer Science, Sri Sathya Sai Institute of Higher Learning, Prasanthi Nilayam, AP 515134 India. email: dhanumantharao@sssihl.edu.in

February 27, 2012 DRAFT

The original publication is available at:
http://www.signalprocessingsociety.org/publications/periodicals/letters/
Speech enhancement using a generic noise codebook

I. INTRODUCTION

Enhancement of speech in the presence of background noise has attracted much interest due to its practical relevance. A particularly challenging application is single-microphone noise reduction in mobile telephony. The low cost of a single-microphone device makes it attractive, but the absence of multiple microphones precludes beamformer-based solutions to suppress high levels of non-stationary noise. A single-microphone approach that works well under non-stationary conditions is thus relevant.

Single-microphone algorithms are also relevant in multi-microphone applications, e.g., hands-free audio and video conferencing systems in reverberant and diffuse non-stationary noise fields, or where there are a number of interfering sources present. Spatial filtering techniques such as beamforming can only achieve limited success in these scenarios and additional noise suppression needs to be performed at the output of the beamformer in a post-processing step, which corresponds to using a single-microphone system.

Knowledge-based speech enhancement methods such as codebook-driven schemes [1] have been shown to perform well under non-stationary noise conditions, even when operating on a single microphone signal. These methods rely on trained codebooks of speech and noise spectral shapes parameterized by e.g., linear predictive (LP) coefficients.

The use of a speech codebook is intuitive and lends itself readily to a practical implementation. In fact they have been successfully employed in speech coding for several years. The speech codebook can either be speaker independent (trained using data from several speakers), or speaker dependent. The latter case is relevant as mobile phones are personal and are often used by a single speaker.

The use of noise codebooks in a practical implementation, however, is challenging due to the variety of noise types encountered in practice. This paper presents a new method that does not require a trained noise codebook. Instead, a generic noise codebook is employed together with a trained speech codebook. Thus, the method is suitable for use in diverse noise environments. Moreover, the generic noise codebook described here can be generated on-the-fly and does not require to be stored offline.
II. SPEECH ENHANCEMENT USING A GENERIC NOISE CODEBOOK

First, the signal model and an overview of the algorithm using trained codebooks of both speech and noise is presented in section II-A. The proposed algorithm that uses a trained speech codebook and a generic noise codebook is described in II-B.

A. Background

Consider an additive noise model where speech and noise are assumed to be independent:

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

where \( y(n), x(n) \) and \( w(n) \) represent the sampled noisy speech, clean speech and noise respectively. Let \( P_y(\omega) \) denote the PSD of the observed noisy signal \( y(n) \). The original codebook-based algorithm described in [1] uses trained codebooks of speech and noise power spectral densities (PSDs) parameterized as LP coefficients. The goal is to use these codebooks to obtain an estimate of the speech and noise PSDs for each short-time segment of noisy speech data. The estimation of the speech and noise PSDs can follow either a maximum-likelihood (ML) approach or a Bayesian minimum mean-squared error (MMSE) approach. The specific choice is not relevant for the description of the proposed method and the ML approach is assumed in the following for ease of exposition.

The relation between a vector of LP coefficients \( \theta_x = (a_{x_0}, \ldots, a_{x_p}) \) and the underlying PSD, where \( a_{x_0} = 1 \) and \( p \) is the LP model order, is given by:

\[ P_x(\omega) = \frac{1}{|A_x(\omega)|^2}, \tag{2} \]

where \( A_x(\omega) = \sum_{k=0}^{p} a_{x_k} e^{-j \omega k} \). A similar relation holds between the noise LP coefficients and the noise PSD \( P_w(\omega) \). Using these relations, the noisy PSD can be modeled as

\[ \hat{P}_y(\omega) = \underbrace{g_x P_x(\omega)}_{\equiv \hat{P}_s(\omega)} + \underbrace{g_w P_w(\omega)}_{\equiv \hat{P}_w(\omega)}, \tag{3} \]

where \( g_x \) and \( g_w \) are the frequency-independent (non-negative) level terms associated with the speech and noise PSDs, which are introduced to account for the variation in the level between the PSDs stored in the codebook and the one encountered in practice.

A search is performed through all pairs of speech and noise codebook entries to determine the pair that maximizes a certain similarity measure between the observed noisy PSD \( P_y(\omega) \) and the modeled PSD \( \hat{P}_y(\omega) \), as described in the following.
Consider a pair of speech and noise PSDs, given by the $i^{th}$ vector from the speech codebook and the $j^{th}$ vector from the noise codebook. The noisy PSD corresponding to this pair can be written as

$$\hat{P}_{ij}^y(\omega) = g_{ij}^x P_x^i(\omega) + g_{ij}^w P_w^j(\omega).$$

(4)

Note that in equation (4) above, $P_x^i(\omega)$ and $P_w^j(\omega)$ are known (given by the respective codebook entries), whereas the level terms $g_{ij}^x$ and $g_{ij}^w$ are unknown. The maximum-likelihood estimate of the desired speech and noise PSDs can be obtained in a two-step procedure. The logarithm of the likelihood that a given pair $g_{ij}^x P_x^i(\omega)$ and $g_{ij}^w P_w^j(\omega)$ resulted in the observed noisy PSD $P_y(\omega)$ is captured by the following equation [2]:

$$L_{ij}(P_y(\omega), \hat{P}_{ij}^y(\omega)) = \int_0^{2\pi} \frac{-P_y(\omega)}{\hat{P}_{ij}^y(\omega)} + \ln \left(\frac{1}{\hat{P}_{ij}^y(\omega)}\right) d\omega$$

$$= \int_0^{2\pi} \frac{-P_y(\omega)}{g_{ij}^x P_x^i(\omega) + g_{ij}^w P_w^j(\omega)} + \ln \left(\frac{1}{g_{ij}^x P_x^i(\omega) + g_{ij}^w P_w^j(\omega)}\right) d\omega.$$  

(5)

In the first step, the unknown level terms $g_{ij}^x$ and $g_{ij}^w$ that maximize $L_{ij}(P_y(\omega), \hat{P}_{ij}^y(\omega))$ are determined. One way to do this is by differentiating (5) with respect to $g_{ij}^x$ and $g_{ij}^w$, setting the result to zero, and solving the resulting set of simultaneous equations. However, these equations are non-linear and not amenable to a closed-form solution. An alternative approach is based on the fact that the likelihood is maximized when $P_y(\omega) = \hat{P}_{ij}^y(\omega)$, and thus the gain terms can be obtained by minimizing the spectral distance between these two entities [2].

Once the level terms are known, the value of $L_{ij}(P_y(\omega), \hat{P}_{ij}^y(\omega))$ can be determined as all entities on the right hand side of (5) are known. This procedure is repeated for all pairs of speech and noise codebook entries to identify the pair that results in the largest likelihood. As this step is performed for every short-time segment, the method can accurately estimate the noise PSD even under non-stationary noise conditions. Let $\{i^*, j^*\}$ denote the pair resulting in the largest likelihood for a given segment, and let $g_{x}^{*}$ and $g_{w}^{*}$ denote the corresponding level terms. Then the speech and noise PSDs are given by

$$\hat{P}_x(\omega) = g_{x}^{*} P_x^{i^*}$$

and

$$\hat{P}_w(\omega) = g_{w}^{*} P_w^{j^*},$$

which can be used to generate a Wiener filter that can be applied to the noisy segment:

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

(7)
B. Using a generic noise codebook

This section discusses the use of a generic noise codebook in the codebook-driven speech enhancement framework. Recall that the method presented in the previous section involves the use of trained speech and noise codebooks. Each codebook entry corresponds to a PSD, and an exhaustive search over the speech and noise codebooks results in an estimate of the speech and noise PSD that maximize the likelihood score. While a trained speech codebook is practical, it is of interest to examine the use of a generic noise codebook so that the method may be applied in diverse noise environments.

Instead of a codebook of trained noise PSDs, a codebook with prototype band-pass flat PSDs is proposed in the new approach. In the simplest case, if the noise codebook has \(N_w\) entries, then the frequency range of interest is divided into \(N_w\) bands, each of a certain width. Each codebook entry corresponds to a flat band-pass PSD, corresponding to the frequency band represented by that entry. The final noise PSD is modeled as a weighted sum of these band-limited flat PSDs, where the weights need to be determined for each short-time segment. The noisy PSD corresponding to the \(i^{th}\) speech codebook entry is thus modeled as

\[
\hat{P}_{yi}(\omega) = g_{xi}^i P_{xi}^{i}(\omega) + \sum_{k=1}^{N_w} g_{xw}^k P_{xw}^k(\omega),
\]

where the gain terms \(g_{x}^i\) and \(g_{xw}^k\), \(1 \leq k \leq N_w\) need to be determined. If the bands are of equal width, the codebook may be defined as

\[
P_{xw}^k(\omega) = \begin{cases} 
1 & \text{for } \omega \in \left[(k-1)\pi/N_w, k\pi/N_w \right], \\
0 & \text{otherwise},
\end{cases}
\]

for \(1 \leq k \leq N_w\) and \(0 \leq \omega \leq \pi\).

Before obtaining the expressions for the unknown gain terms, some remarks are in order regarding the approach of modeling the noise PSD as a weighted sum of band-limited flat PSDs. First, the weighted sum approach is able to model colored noise, where the frequency resolution is determined by the width of each band, which in turn is determined by the number of codebook entries \(N_w\). Secondly, one entry of the noise codebook can be dedicated to storing the most recent estimate of the noise PSD obtained from one of the state-of-the-art noise estimation algorithms, e.g., [3]. In this manner, the algorithm may be expected to perform at least as well as the existing algorithms, and perform better under difficult conditions. Finally, it may be noted that in contrast to the method of Section II-A that involved a search across all combinations of speech and noise codebook entries, when using the generic noise codebook, it is sufficient to loop once over all the speech codebook entries.
It is important to note that while the frequency resolution increases as $N_w$ increases, it is not possible to employ a noise codebook with the same frequency resolution as the speech codebook. The reason is that the freedom allowed by the gain terms $g_{k}^i$ in (8) implies that these gain terms can be adjusted such that every speech codebook entry results in an equally high likelihood in (5). A coarse frequency resolution (having a common gain term for a band of frequency bins) in the noise codebook ensures that the speech codebook entries that result in a large likelihood are also close to the underlying clean speech PSD.

The gain terms $g_{i}^i$ and $g_{w}^k$ can be determined as those that maximize $L_i(P_y(\omega), \hat{P}_y(\omega))$, which from (5) is given by:

$$L_i(P_y(\omega), \hat{P}_y(\omega)) = \int_0^{2\pi} -\frac{P_y(\omega)}{g_{x}^iP_x^i(\omega) + \sum_{k=1}^{N_w} g_{w}^k P_k^i(\omega) + \hat{P}^i_{\text{est}}P_{\text{est}}} d\omega,$$

where $P_{\text{est}}^i$ is the estimate of the noise PSD obtained from a conventional noise estimation scheme, e.g., [3], and $g_{\text{est}}^i$ is its corresponding gain term. As explained in Section II-A, this is done by minimizing the spectral distance between $P_y(\omega)$ and $\hat{P}_y(\omega)$. First, for notational convenience, the speech and noise PSDs and the gain terms are renamed as follows:

$$P_1(\omega) = P_x^i(\omega), P_2(\omega) = P_{w}^i(\omega), \ldots, P_{N_w+1}(\omega) = P_{w}^{N_w}(\omega), P_{N_w+2}(\omega) = P_{\text{est}}(\omega)$$

$$g_1 = g_x^i, g_2 = g_{w}^1, \ldots, g_{N_w+1} = g_{w}^{N_w}, g_{N_w+2} = g_{\text{est}},$$

so that

$$\hat{P}_y(\omega) = \sum_{k=1}^{N_w+2} g_k P_k(\omega).$$

The cost function to be minimized is

$$\xi = \int_0^{2\pi} \left( P_y(\omega) - \sum_{k=1}^{N_w+2} g_k P_k(\omega) \right)^2 d\omega,$$

the partial derivative of which with respect to $g_l$, $1 \leq l \leq N_w + 2$ can be set to zero to solve for the gain terms:

$$0 = \frac{d}{dg_l} \xi = \int_0^{2\pi} \left( P_y(\omega) - \sum_{k=1}^{N_w+2} g_k P_k(\omega) \right) P_l(\omega) d\omega, \quad 1 \leq l \leq N_w + 2.$$
where

\[
A = [a_{kl}] \quad 1 \leq k, l \leq N_w + 2,
\]

\[
a_{kl} = \int_0^{2\pi} P_k(\omega)P_l(\omega) d\omega,
\]

\[
g = [g_1, g_2, \ldots, g_{N_w+2}]^T,
\]

\[
b = \left[ \int_0^{2\pi} P_y(\omega)P_l(\omega) \right]_{1 \leq l \leq N_w+2}.
\]

Directly solving the unconstrained problem in (14) is not guaranteed to result in non-negative gain terms. Instead, a few sub-systems of the linear system (14) are solved, and among those sub-systems with non-negative gain terms, the one resulting in the highest likelihood is used as the solution. The considered sub-systems correspond to different possible scenarios:

1) The conventional noise estimate provides the best description of the noisy PSD, e.g., in a stationary noise-only region. Here, \( \hat{P}_y(i\omega) = g_{N_w+2}P_{N_w+2}(\omega) \) (see (12)). The linear system (14) reduces to

\[
A = a_{N_w+2}N_w+2, g = g_{N_w+2}, b = b_{N_w+2}. The other gain terms, \( g_l, 1 \leq l < N_w + 2 \), are set to zero.
\]

2) The generic noise codebook and the conventional noise estimate provide the best description of the noisy PSD, e.g., in a non-stationary noise-only region. Here, \( \hat{P}_y(i\omega) = \sum_{k=2}^{N_w+2} g_kP_k(\omega) \). The linear system (14) reduces to

\[
A = a_{kl}, g = g_l, b = b_l, 2 \leq k, l \leq N_w + 2. g_1 is set to zero.
\]

3) Speech only. Here, \( \hat{P}_y(i\omega) = g_1P_1(\omega) \) so that

\[
A = a_{11}, g = g_1, b = b_1. The other gain terms, \( g_l, 1 \leq l < N_w + 2 \), are set to zero.
\]

4) Speech + conventional noise estimate. Here, \( \hat{P}_y(i\omega) = g_1P_1(\omega) + g_{N_w+2}P_{N_w+2}(\omega) \) so that

\[
A = \begin{bmatrix}
a_{11} & a_{1N_w+2} \\
a_{1N_w+2} & a_{N_w+2N_w+2}
\end{bmatrix}, and b and g are defined analogously. The other gain terms, \( g_l, 2 \leq l < N_w + 1 \), are set to zero.
\]

5) Speech + generic noise codebook + conventional noise estimate, which corresponds to the complete system in (14).

The linear system corresponding to each of the above five cases is solved. Solutions resulting in negative gain terms are discarded, and the one with the highest likelihood score among the valid solutions is considered.

III. EXPERIMENTS

The performance of codebook-based speech enhancement using the proposed generic noise codebook scheme was compared against using a trained noise codebook. The comparison was performed under two
Fig. 1. A segment of the traffic noise used in the experiments.

conditions, one where the noise type used to train the noise codebook was the same as the noise type in the noisy test files (matched scenario for the trained noise codebook based method), and the other where the noise type during training and testing was different (mismatched scenario for the trained noise codebook based method). Note that in the matched scenario, only the type was similar across training and testing, and the actual samples used were different. The Bayesian MMSE approach described in [1], together with the new gain estimation technique necessary when using a generic noise codebook, was used for the estimation of the speech and noise PSDs to construct the Wiener filter.

Noisy speech was generated by adding traffic noise to ten clean speech utterances from the Wall Street Journal (WSJ) speech database [4] at a sampling frequency of 8 kHz and a signal-to-noise ratio (SNR) of 0 dB. The traffic noise was highly non-stationary, corresponding to the noise of fast moving vehicles on a busy road observed by a pedestrian standing at a fixed point, and a segment is shown in Fig. 1. A database of 180 distinct training utterances of duration around 5 sec each from a single speaker were used to train a 256-entry speaker dependent codebook of tenth order LP coefficients using the LBG algorithm [5] and the root mean squared log-spectral distortion (LSD) as the error criterion. The LP coefficients were extracted from Hann-windowed segments of length 256 samples with a 50% overlap. A four-entry codebook of sixth order LP coefficients trained on traffic noise was used in the best-case scenario and an four-entry codebook of sixth order LP coefficients trained on cafeteria noise was used in the worst-case scenario for the trained noise codebook-based method. The generic noise codebook consisted of four non-overlapping bands, generated according to (9).

The performance of the two schemes was quantified using segmental SNR (SSNR) and log-likelihood ratio (LLR) as the objective quality measures, averaged over the ten utterances. SSNR is conventionally used to report noise reduction performance, and we include LLR as it has been shown to have a high correlation to subjective quality [6]. Note that the reported values are the improvement with respect to
the noisy input, and thus larger values indicate better performance.

<table>
<thead>
<tr>
<th></th>
<th>Generic noise CB</th>
<th>Trained noise CB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improvement in SSNR</td>
<td>6.3 dB</td>
<td>7.2 dB</td>
</tr>
<tr>
<td>Improvement in LLR</td>
<td>0.14</td>
<td>0.15</td>
</tr>
</tbody>
</table>

**TABLE I**
RESULTS FOR THE MATCHED SCENARIO FOR THE TRAINED NOISE CODEBOOK.

Table I shows the results for the matched scenario. In terms of SSNR, using the trained noise codebook clearly results in better performance than the generic noise codebook. This result is intuitive and demonstrates the strength of a trained codebook in a matched condition. The LLR results are similar for both methods. The benefit of using a generic noise codebook becomes apparent in the mismatched scenario, the results of which are shown in Table II. The trained codebook now results in worse performance than the generic codebook, with a difference of over 2 dB in terms of SSNR, and a noticeable difference in LLR. Note that in the mismatched case the trained noise codebook was trained on cafeteria noise whereas the noisy speech contained traffic noise. The SSNR and LLR values for the generic noise codebook are the same in both scenarios as the noisy test data and generic noise codebook remain the same. Based on these results one may conclude that while there is a penalty (0.9 dB in SSNR) to be paid when using a generic codebook in a matched scenario, the significant gain (2.1 dB in SSNR) in terms of robustness in a mismatched scenario justifies its use over a trained codebook. Moreover, the generic noise codebook approach does not require the training and storage of multiple noise codebooks to account for diverse noise environments.

<table>
<thead>
<tr>
<th></th>
<th>Generic noise CB</th>
<th>Trained noise CB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improvement in SSNR</td>
<td>6.3 dB</td>
<td>4.2 dB</td>
</tr>
<tr>
<td>Improvement in LLR</td>
<td>0.14</td>
<td>0.09</td>
</tr>
</tbody>
</table>

**TABLE II**
RESULTS FOR THE MISMATCHED SCENARIO FOR THE TRAINED NOISE CODEBOOK.
IV. CONCLUSIONS

In this letter, the use of a generic noise codebook has been proposed for codebook-based speech enhancement algorithms. Such a codebook does not require the training of noise codebooks for different noise types and can be generated on-the-fly, thereby enabling a practical deployment of codebook-based algorithms. Moreover, it does not suffer from the disadvantage of trained noise codebooks in the event of a mismatch between training and testing conditions, as confirmed by experimental results.

REFERENCES


