interspeech 2008 incorporating SST 08

Speech Enhancement based on Hypothesized Wiener Filtering

V. Ramasubramanian, Deepak Vijaywargi†

Siemens Corporate Technology - India, Bangalore, India

V.Ramasubramanian@siemens.com, deepak@ee.washington.edu

Abstract

We propose a novel speech enhancement technique based on the hypothesized Wiener filter (HWF) methodology. The proposed HWF algorithm selects a filter for enhancing the input noisy signal by first 'hypothesizing' a set of filters and then choosing the most appropriate one for the actual filtering. We show that the proposed HWF can intrinsically offer superior performance to conventional Wiener filtering (CWF) algorithms, which typically perform a selection of a filter based only on the noisy input signal which results in a sub-optimal choice of the filter. We present results showing the advantages of HWF based speech enhancement over CWF, particularly with respect to the baseline performances achievable by HWF and with respect to the type of clean frames used, namely, codebooks vs a large number of clean frames. We show the consistently better performance of HWF based speech enhancement (over CWF) in terms of spectral distortion at various input SNR levels.

Index terms: Speech enhancement, hypothesized Wiener filtering, iterative Wiener filtering

1. Introduction

Speech enhancement is a fundamental problem in speech processing and has received significant attention from researchers owing to its importance in communication, robustness of speech- and speaker-recognition systems etc.. The problem of enhancing speech corrupted with additive background noise is a particularly important problem where a communications system or an automatic speech- or speaker-recognition system has to operate in noisy environments.

A conventional approach to dealing with noisy speech by speech enhancement (for applications such as speech recognition, speaker-recognition, speech coding etc.) is to apply noise-removal techniques such as spectral-subtraction [1], [2] or conventional Wiener filtering methods [3] so as to get an enhanced speech signal prior to feature extraction. While spectral subtraction requires an estimate of the noise power spectral densities, typically from the most recent non-speech region, Wiener filtering methods require estimates of both clean speech power spectrum and the noise power spectrum. There are a wide variety of Wiener filtering techniques depending on how the clean speech power spectrum estimate is obtained for any given frame; these can be broadly categorized as based on spectral-subtraction or, estimates of signal spectrum from previous 'cleaned' frames or, from model-based estimates such as linear-prediction, or using vector quantizer codebooks; these methods are typically employed in an iterative framework [4], [3].

In this paper, we propose a novel speech enhancement technique based on the hypothesized Wiener filter (HWF) methodology. The HWF framework was first proposed by [5] for robust speaker-dependent isolated word recognition in a DTW framework and has subsequently been adapted to HMM frameworks using state-based filtering for noisy speech recognition. More recently, we used the HWF framework for a high performance noise-robust text-dependent speaker recognition [6].

However, despite its appealing feature of making use of clean templates or HMMs, it has not been used for speech enhancement so far. In this paper, we propose the basic formulation of HWF for speech enhancement and show how the proposed HWF can intrinsically offer superior performance to conventional Wiener filtering (CWF) algorithms, specifically within the codebook constrained iterative Wiener filtering framework as proposed earlier in [4]. We show that CWF algorithms select a filter for enhancing the input noisy signal in a sub-optimal way; in contrast, we show that the proposed HWF formulation is more optimal since it 'first hypothesizes' a set of Wiener filters and only then chooses the most appropriate one for the actual filtering. We present results showing the performance advantages of HWF based speech enhancement over CWF, particularly with respect to the optimal baseline performances achievable by HWF and also with respect to the nature of clean frames used, namely, codebooks vs a large number of unclustered clean frames. We show the consistently superior performance of HWF based speech enhancement in terms of spectral distortions.

2. Conventional Wiener filtering (CWF)

Here we describe briefly the essential formulation of conventional Wiener filter (CWF) algorithms set in a codebook constrained manner [4] in a non-iterative framework in order to provide the contrasting basis for the HWF methodology proposed here.

The noisy input speech is represented as a sequence of T frames $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_T$, where $\mathbf{x}_i$ is the sequence of noisy speech samples of the i^{th} frame. Let the power spectral density (psd) of the noisy input speech be $P_{\mathbf{x}_1}(w)$, $P_{\mathbf{x}_2}(w)$, $\ldots$, $P_{\mathbf{x}_i}(w)$, $\ldots$, $P_{\mathbf{x}_T}(w)$ and the corresponding sequence of MFCC feature vectors be $X_1, X_2, \ldots, X_i, \ldots, X_T$.

In conventional CWF formulations such as in [4], this noisy speech is enhanced by Wiener filtering using a codebook of N clean frames $\mathcal{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_j, \ldots, \mathbf{y}_N)$. These are typically in the form of the LSF vectors derived by a vector quantization codebook design algorithm applied on a large number of clean LSF vectors. Here, for ease of generalization, we assume that $\mathcal{Y}$ above is a set of clean frames (of time domain signal). Let the psd of the set of these codebook of N clean frames be $P_{\mathbf{y}_1}(w)$, $P_{\mathbf{y}_2}(w)$, $\ldots$, $P_{\mathbf{y}_j}(w)$, $\ldots$, $P_{\mathbf{y}_N}(w)$ and let the corresponding MFCC feature vectors be $Y_1, Y_2, \ldots, Y_j, \ldots, Y_N$. Let $P_n(w)$ be the noise-estimate obtained from the most recent non-speech region of the input noisy speech.

In the basic Wiener filtering operation in the CWF formulations, the noisy time domain signal frame $\mathbf{x}_i$ is filtered

†Presently with University of Washington, Seattle

10.21437/Interspeech.2008-39

by a Wiener filter derived by the following steps to yield the 'cleaned' signal $\tilde{\mathbf{x}}_{i,j(i)}$:

1. Choose best clean frame $\mathbf{y}_{j(i)}$: The noisy input frame $\mathbf{x}_i$ is vector quantized using the clean codebook $\mathcal{Y}$ to obtain the 'best' clean frame $\mathbf{y}_{j(i)}$ as:

$$j(i) = \arg \min_{k=1,\ldots,N} d(i,k) \tag{1}$$

where $d(i,k)$ is given by

$$d(i,k) = d(X_i, Y_k) \tag{2}$$

Note that we have defined the distance used in vector quantization as based on the Euclidean distance between the MFCC vectors corresponding to the noisy frame $\mathbf{x}_i$ and a clean frame $\mathbf{y}_j$ for the sake of consistency with the later description of our proposed HWF formulation. This could as well be based on the distance between the LSF vectors of these frames, as is usually reported in earlier work [4], without any loss of theoretical accuracy or actual performance.

2. Define Wiener filter of best clean frame: The Wiener filter derived by using this 'best' clean frame $y_{j(i)}$ is given by,

$$W_{j(i)}(w) = \frac{P_{\mathbf{y}_{j(i)}}(w)}{P_{\mathbf{y}_{j(i)}}(w) + P_n(w)} \tag{3}$$

3. Wiener filter noisy frame: The filtering of the signal at the i^{th} frame by the $j(i)^{th}$ clean frame is done by computing the 'cleaned' frame psd as,

$$P_{\tilde{\mathbf{x}}_{i,j(i)}}(w) = P_{\mathbf{x}_i}(w) \cdot W_{j(i)}(w) \tag{4}$$

and then in obtaining $\tilde{\mathbf{x}}_{i,j(i)}$ as a frame of time domain samples corresponding to the psd $P_{\tilde{\mathbf{x}}_{i,j(i)}}(w)$.

The critical issue to note is that the best clean psd for the conventional Wiener filtering (CWF) operation is determined based on how close a clean frame is to the 'noisy' input frame. However, this choice is bound to be a bad one given that the noisy frame, by itself, has no means of selecting the best clean frame to which the clean frame (underlying the noisy frame) might be close to, i.e., the fact that the input frame is noisy will lead the vector quantization based determination of the 'best' clean frame to be error prone in the sense that this 'best' clean frame will not in any way be the 'best' estimate of the 'clean' frame underlying the 'noisy' frame for Wiener filtering of the noisy frame. Hence the use of such an estimate of the clean speech will lead to a sub-optimal Wiener filtering in CWF.

3. Proposed HWF based speech enhancement

In contrast to the above CWF approach (for selecting the 'best' clean frame used in the Wiener filtering of the noisy frame), the proposed hypothesized Wiener filtering (HWF) uses a measure of merit for selecting the 'best' clean frame 'after' Wiener filtering the input noisy frame with 'every possible' clean frame (available as in a clean codebook); i.e, in HWF, we follow the following steps which includes as a first step a 'hypothesized Wiener filtering' operation, in contrast to the CWF algorithm as above:

1. Hypothesized Wiener filtering (HWF): 'Every' clean frame is used for Wiener filtering the input noisy frame to yield N candidate 'cleaned' frames.

2. HWF distance: We then define the distance between the 'cleaned' frame and the clean frame used to obtain the 'cleaned' frame and select the 'best' clean frame as the clean frame which yields the minimum of this distance.

3. Wiener filtering using best hypothesized Wiener filter (HWF): This 'best' clean frame defines the 'optimal' Wiener filter for the 'actual' Wiener filtering of the input noisy frame, with the corresponding 'cleaned' frame as the enhanced speech output.

Now, we describe the proposed HWF formulation in detail.

1. Hypothesized Wiener filtering: Wiener filter $\mathbf{x}_i$ by all clean frames $\mathcal{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_j, \ldots, \mathbf{y}_N)$; i.e., for each clean frame $\mathbf{y}_j, j = 1, \ldots, N$, obtain corresponding Wiener filter

$$W_j(w) = \frac{P_{\mathbf{y}_j}(w)}{P_{\mathbf{y}_j}(w) + P_n(w)} \tag{5}$$

Wiener filter noisy input frame $\mathbf{x}_i$ using $W_j(w)$ by

$$P_{\tilde{\mathbf{x}}_{i,j}}(w) = P_{\mathbf{x}_i}(w) \cdot W_j(w) \tag{6}$$

and obtain $\tilde{\mathbf{x}}_{i,j}$ as a frame of time domain samples corresponding to the psd $P_{\tilde{\mathbf{x}}_{i,j}}(w)$.

2. Choose best clean frame $\mathbf{y}_{j(i)}$: Obtain $\tilde{X}_{i,j}$ as MFCC vector of the cleaned speech signal of the i^{th} frame given by $\tilde{\mathbf{x}}_{i,j}$ obtained as above in Step 1 for all $j = 1, \ldots, N$. Define the 'HWF distance' between the noisy i^{th} frame and the j^{th} clean frame as

$$d_w(i,j) = d(\tilde{X}_{i,j}, Y_j) \tag{7}$$

where $d(\tilde{X}_{i,j}, Y_j)$ is the Euclidean distance between the MFCC vectors $\tilde{X}_{i,j}$ and Y_j (Y_j being the MFCC vector of the j^{th} clean frame). By this, a set of HWF distances $d_w(i,j), j = 1, \ldots, N$ is obtained for the i^{th} input frame, which contains the distances between the 'clean' frames $\mathbf{y}_j, j = 1, \ldots, N$ and the corresponding 'cleaned' frames $\tilde{\mathbf{x}}_{i,j}, j = 1, \ldots, N$.

The best clean frame $\mathbf{y}_{j(i)}$ that should be used to define the Wiener filter (for filtering the noisy frame i) is obtained as that j which minimizes the HWF distance in the set of distances $d_w(i,j), j = 1, \ldots, N$ for input noisy frame i, as given by

$$j(i) = \arg \min_{k=1,\ldots,N} d_w(i,k) \tag{8}$$

Subsequent to this choice of the 'best' clean frame $j(i)$ for the i^{th} noisy frame, we obtain the enhanced frame of the i^{th} frame as $\tilde{\mathbf{x}}_{i,j(i)}$ from $P_{\tilde{\mathbf{x}}_{i,j(i)}}(w)$ in Eqn. 3.

The central ideas behind the above HWF based Wiener filtering is as follows:

1. For large N, and if the N 'clean' frames are ensured to come from a large set of clean speech data representative of all possible phonetic content (even across a large collection of languages, speakers, environmental characteristics etc), then one of the 'clean' frames in the 'clean codebook' is bound to be a good approximation of the 'clean' version of a frame of the input noisy speech.
2. In the proposed HWF formulation, this 'best' clean frame is determined by the lowest $d_w(i,j)$ from the set of N HWF distances for any i, where the corresponding $j^* = j(i)$, i.e., some clean frame $\mathbf{y}_j(i)$) which (when used as the clean psd estimate in the Wiener filter) performs the 'best' Wiener filtering of the noisy input frame

The key concept here is that HWF finds this 'best' Wiener filtering through the HWF distance which represents how close the 'cleaned' signal $\mathbf{x}_{i,j(i)}$ has become to the 'clean' signal (unknown) underlying the 'noisy' signal of frame i. This in turn is measured as the (HWF) distance between the 'cleaned' signal and the 'known clean' frame $\mathbf{y}_{j(i)}$ which is now after all the best estimate of the unknown 'clean' signal underlying the noisy frame (as used in the Wiener filtering operation leading to the 'cleaned' signal).

Clearly, this HWF based Wiener filtering is more optimal than CWF due to three important differences: i) All candidate

clean frames perform Wiener filtering before selection in HWF, while in CWF the clean frame is selected 'prior' to Wiener filtering, ii) the distance used in CWF is between the noisy frame and the clean frame, and does not represent how well the clean frame would have been effective (as the clean psd estimate in the Wiener filter) in cleaning (i.e., Wiener filtering) the noisy frame towards making it closer to the underlying clean frame, iii) the distance used in our proposed HWF is more represntative of how close the 'cleaned' signal is to the 'clean' signal, which after all is the real measure of how effective a clean frame (used in defining the Wiener filter operation) is.

4. Iterative forms of CWF and HWF

Here we consider CWF set in an iterative framework, particularly the codebook constrained iterative Wiener filtering algorithms [3], [4] and bring out the essential differences between these and the proposed HWF based speech enhancement algorithm in iterative form.

In a typical iterative conventional Wiener filter (ICWF) framework, the Wiener filtering is done successively using the cleaned pdf of the input signal. This is given as follows. The Wiener filter for any iteration r is given by,

$$W_r(w) = \frac{P_{\mathbf{y}_{j^r}(w)}}{P_{\mathbf{y}_{j^r}(w)+P_{n(r)}(w)}} \tag{9}$$

and the Wiener filter iteration is given by

$$P_{\tilde{\mathbf{x}}_i(r+1)} = P_{\tilde{\mathbf{x}}_i(r)} \cdot W_r(w) \tag{10}$$

where $\mathbf{y}_{j^r}$ is the clean frame whose psd was chosen as the best (for Wiener filtering the input noisy signal at iteration r, i.e., $\tilde{\mathbf{x}}_i(r)$), based on the closeness between the corresponding MFCC vectors $\tilde{X}_i(r)$ and $Y_j(r)$ as given by

$$j^r = \arg \min_{k=1,\ldots,N} d(\tilde{X}_i(r), Y_k) \tag{11}$$

$P_{n(r)}$ being the noise-estimate from the cleaned signal at iteration r.

While the above equations give the iterative Wiener filter formalism in the ICWF, the corresponding iterative hypothesized Wiener filter (IHWF) framework has similar iterative structure except for the difference in the definition of the clean frame $\mathbf{y}_{j^r}$ used in the Wiener filtering operation in Eqn. (9) whose index j^r is given by,

$$j^r = \arg \min_{k=1,\ldots,N} d(\tilde{X}_{i,k}(r+1), Y_k) \tag{12}$$

where $\tilde{X}_{i,k}(r+1)$ is the MFCC vector derived from the 'cleaned' signal psd $P_{\tilde{\mathbf{x}}_{i,k}(r+1)}(w)$ obtained as $P_{\tilde{\mathbf{x}}_{i,k}(r+1)}(w) = P_{\mathbf{x}_i(r)}(w) \cdot W_k(w)$, i.e., by Wiener filtering the noisy frame $\tilde{\mathbf{x}}_i(r)$ (at iteration r) using the Wiener filter corresponding to a clean frame $\mathbf{y}_k$ as given in Eqn. (5). By this, the HWF chooses the best filter from the clean frames at every iteration based on an 'hypothesis' of all N filters, in the same way as is done in the non-iterative formalism in Sec. 3, but now embedded within any given iteration.

5. Experiments and results

Here, we show the results of speech enhancement using CWF and HWF both in non-iterative form and iterative form. Moreover, we consider the use of a clean codebook vs a large set of clean frames and indicate the advantages of the latter in realizing performances closer to an empirical baseline which the HWF can offer. In order to provide a comparison between HWF and CWF, we use spectral distortion between the enhanced signal and the original signal as an objective measure of the effectiveness of the two algorithms. All the experiments are done using the TIMIT database, with white noise added from the NOISEX92 database with SNRs 0, 5 and 10 dB.

First we focus on the differences between CWF and HWF in a codebook constrained Wiener filtering framework as proposed earlier in [4]. Here we construct clean codebooks of size 1 to 1024 in powers of 2 from a large set of clean frames. The clean frames in the above formalism of HWF are now these codebooks whose size N varies from 2 to 1024.

Fig. 1 shows the spectral distortion (SD in dB) between CWF and HWF for codebook sizes 1 to 1024 on a test sentence of 3 seconds long taken from TIMIT database. It can be clearly observed that HWF offers a significant performance advantage over CWF in all the codebook sizes. Moreover, we also show the performance baseline achievable by HWF, when the clean frames are the same as the clean frames underlying the noisy input sentence. By this, the HWF has the choice of selecting the identical clean frame for cleaning the corresponding noisy frame, and hence provides the best performance achievable (though the actual clean frame may not get selected and the SD is not 0 dB due to mismatch in noise-estimate and actual noise spectrum and mismatch between MFCC distance for selection (Eqn. (7)) and SD performance measure). Clearly, it can be seen that even a codebook of size 1024 does not offer sufficient spectral resolution or accuracy to reach this baseline.

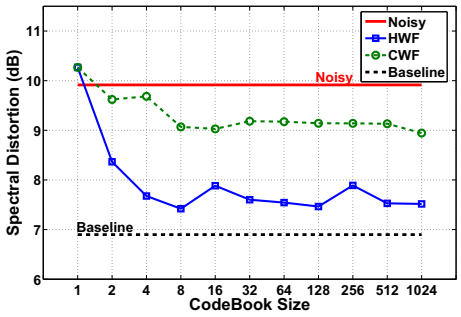

Figure 1: *Comparison of conventional Wiener filtering (CWF) and proposed hypothesized Wiener filtering (HWF) using clean codebooks (for 0 dB noisy speech)*

We now show the performance of the codebook constrained iterative frameworks for a codebook of size 256 in Fig. 2. Due to the gross error in selecting a poor clean frame by the ICWF algorithm in the first iteration (as shown in Fig. 1), the successive iterations of ICWF does not perform adequately. In fact, it can be observed that the IHWF in the first iteration offers significantly better performance not only for the first iteration of ICWF, but also over all successive iterations of ICWF and that ICWF performance is only marginally better that its own first iteration performance and considerably poorer than the IHWF performance or the IHWF baseline performance.

Now, we turn our attention to the crucial question of why it is necessary to use a small codebook at all, when it is of more importance to be able to reach the baseline performance when the clean frames of the noisy input speech is used. Since this is not possible in practical conditions, it is only possible to approximate this with a large number of clean frames. We do this by using the raw clean frames from a large number of TIMIT sentences, excluding the test sentence and the test speaker. This is a realistic speech enhancement scenario where the system has to be equipped to deal with input speech of any phonetic content,

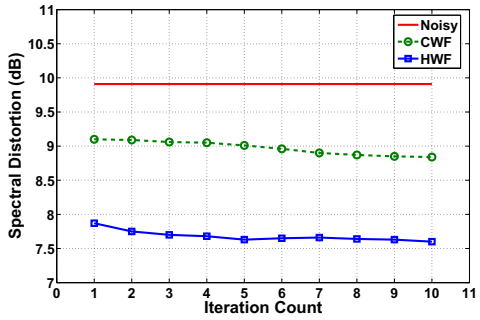

Figure 2: *Comparison of iterative forms of conventional Wiener filtering (ICWF) and proposed hypothesized Wiener filtering (IHWF) using clean codebooks (for 0 dB noisy speech)*

speaker, language etc. and any intrinsic feature of the algorithm (such as the set of clean frames used) be good for all such test conditions.

Fig. 3 shows the performance of HWF for a test sentence with the corresponding clean frames drawn varying from 10 to number of frames in the entire sentence for 3 cases: i) A: The same sentence but spoken by another speaker, ii) B: A different sentence spoken by the same speaker and iii) C: A different sentence spoken by a different speaker. It can be noted that while A offers the best performance due to match of phonetic content, cases B and C perform with an higher spectral distortion. However, in keeping with the requirement that the system be equipped with clean frames to cater to a wide variety of input utterances, we continue the experiment C, using much larger number of clean frames, i.e., drawn from up to 30 sentences. This is shown in Fig. 4. It can be seen that the HWF performance now reaches the baseline performance quite closely, and that CWF performs poorly.

In comparing Fig. 1, Fig. 3 and Fig. 4, the advantage of using a large number of frames rather than small codebooks can be clearly noted. This advantage arises mainly from the fact that the use of a large number of clean frames has higher scope for selecting a clean frame that is a good approximation of the clean speech underlying the noisy input utterance, as pointed out earlier.

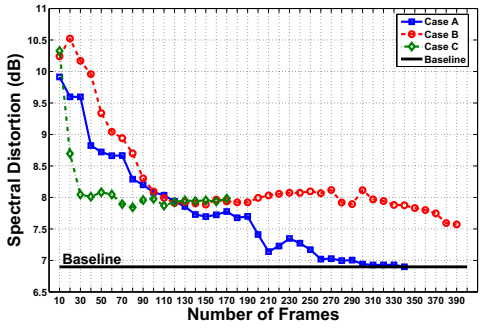

Figure 3: *Performance of proposed hypothesized Wiener filtering (HWF) using varying number of clean frames from different sentences A, B and C (for 0dB)*

In a summarizing result, we show in Fig. 5 the performances of CWF and HWF (along with the noisy performance) in terms of spectral distortion (dB) for various SNRs of 0, 5 and 10 dB for number of clean frames of size about 9000 frames. The HWF can be clearly noted to offer a significantly better performance over CWF in all the cases. In all the experiments reported here, we have consistently observed the HWF based enhancement to be considerably superior to CWF with high intelligibility and very low artefactual noises due to the enhancement process.

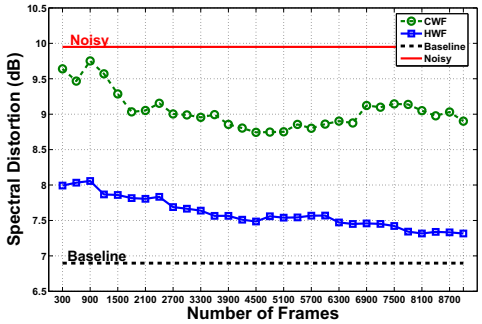

Figure 4: *Comparison of conventional Wiener filtering (CWF) and proposed hypothesized Wiener filtering (HWF) using a large number of clean frames from up to 30 sentences (for 0dB)*

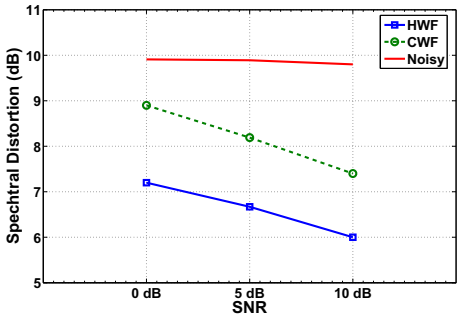

Figure 5: *Comparison of CWF and proposed HWF using a large number of clean frames (approx 9000) for SNRs of 0, 5, 10 dB*

6. Conclusions

We have proposed a novel speech enhancement technique based on the hypothesized Wiener filter (HWF) methodology which selects a filter for enhancing the input noisy signal by first 'hypothesizing' a set of filters and then choosing the most appropriate one for the actual filtering. We present results showing the advantages of HWF based speech enhancement over conventional Wiener filtering (CWF), with respect to the baseline performances achievable by HWF and with respect to the type and number of clean frames used and show the consistently better performance of HWF based speech enhancement (over CWF) in terms of spectral distortion.

7. References

[1] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on Acoustics, Speech and Signal Proc.*, 27(2):113–120, Apr 1979.

[2] M. Berouti, R. Schwartz and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Proc. ICASSP'79*, pages 208–211, 1979.

[3] T. F. Quatieri. *Discrete-time speech signal processing – Principles and practice*. Pearson Education, 2002.

[4] T. V. Sreenivas and P. Kirnapure. Codebook constrained Wiener filtering for speech enhancement. *IEEE Trans. on Acoust., Speech and Sig. Proc.*, 4(5):383-389, Sep 1996.

[5] A. D. Berstein and I. D. Shallom. An hypothesized wiener filtering approach to noisy speech recognition. In *Proc. ICASSP'91*, pages 913–916, 1991.

[6] V. Ramasubramanian, Deepak Vijaywargiay, V. Praveen Kumar. Highly noise robust text-dependent speaker recognition based on hypothesized Wiener filtering. *In Proc. of Interspeech 2006*, pages 1455-1458, Pittsburgh, Sept. 2006.