INVENTORY BASED SPEECH ENHANCEMENT FOR SPEAKER DEDICATED SPEECH COMMUNICATION SYSTEMS

Xiaoqiang Xiao and Peng Lee*

Department of Electrical Engineering
The Pennsylvania State University
University Park, PA 16802
xxx106@psu.edu*

Robert M. Nickel

Department of Electrical Engineering
Bucknell University
Lewisburg, PA 17837
robert.nickel@bucknell.edu

ABSTRACT

We are presenting a method for the enhancement of speech in speaker dedicated speech communication systems. The proposed procedure is fundamentally different from most state-of-the-art filtering approaches. Instead of filtering a distorted signal we are re-synthesizing a new "clean" signal based on its likely characteristics. These characteristics are estimated from the distorted signal. We present a successful implementation of the proposed method for a communication system for which speaker enrollment and noise enrollment are feasible. Forty minutes of clean speech training data is usually sufficient for successful denoising. The proposed method compares very favorably to other state-of-the-art systems in both objective and subjective speech quality assessments.

Index Terms – Speech Enhancement, Hidden Markov Models, Harmonic Tunnelling, Sinusoidal Speech Model, Speaker Dependent Denoising.

1. INTRODUCTION

The distortion of speech signals with additive noise is one of the most hampering factors in speech signal processing today. Human listeners are usually able to (psycho-acoustically) reject high levels of background noise. In contrast, mild levels of noise can interfere significantly in automatic speech recognition and speech coding [1].

The various modern approaches to denoising of speech are mostly *waveform filtering* based methods. Waveform filtering implies that only limited assumptions are made about the specific nature of the underlying signal (i.e. than that it is an acoustic waveform). The most prominent examples of waveform processing are the Wiener filtering extensions proposed by McAulay and Malpass in 1980 [1] and Ephraim and Malah in 1984 [1]. Other examples include schemes that employ wavelets [2] and modifications of the iterative Wiener filter and the Kalman filter [3]. A powerful method in the presence of speech babble noise is the multiband spectral subtraction method proposed by Kamath and Loizou in 2002 [1].

More recently, model based denoising methods have been proposed [4]. In model based denoising a deterministic or stochastic parametric model for a speech signal (and its properties) is used instead of a general waveform model. A popular choice for a speech model in this context is the harmonic plus noise model (HNM) which was studied by Zavarehei, Vaseghi, and Yan [5]. Accurate modeling and estimation of speech and noise gains via hidden Markov models was proposed by Zhao and Kleijn [6]. Codebooks of linear predictive coefficients and their employment for speech denoising within a maximum-likelihood framework was studied by Srinivasan, Samuelsson, and Kleijn [7].

The model based speech denoising method proposed in this paper is inspired by the increasing success of *inventory based speech synthesis systems* [8]. We are assuming that speaker enrollment and noise enrollment are feasible for the given denoising task. The speaker enrollment procedure provides us with training data that can be appropriately clustered and used as an inventory for a "clean" speech signal model. The inventory is augmented with a statistical analysis of the speech signal under clean and noisy conditions. The details of the proposed method are summarized in section 2. Experimental results and performance studies are provided in section 3.

Applications for the proposed method include vehicular speakerphone communication systems and jet pilot communication systems.

2. METHODS

A block diagram of the proposed method is shown in figure 1. The enhancement procedure is divided into three main tasks: (A) a system training task (dashed arrows in figure 1), (B) the signal preprocessing task (dotted arrows), and (C) the signal denoising task (solid arrows).

The system training task consists of the development of a *speech waveform inventory*, two *mel-frequency cepstral coefficient* (MFCC) codebooks (under clean and noisy conditions), and a *hidden Markov model* (HMM) to model the codeword transition statistics under clean and noisy conditions. The details of the system training task are summarized in section 2.1.

The procedures of the signal preprocessing task are adjusted according to the expected noise type. No preprocessing is necessary in the case of white noise. Stationary colored noise requires preprocessing with a *prewhitening filter.* Nonstationary noise is preprocessed with a combination of an estimation of the *power spectral density* of the noise (via *harmonic tunnelling*) and subsequent *Wiener filtering.* The details of the signal preprocessing task are summarized in section 2.2.

Lastly, the speech denoising task combines the results of the preprocessing with the results of a *state sequence computation* from the trained HMM (described in section 2.1). Suitable sections from the speech inventory are chosen through an *inventory unit selection scheme* and are then concatenated to form the targeted denoised speech signal.

Before we discuss the three main components of the proposed method in detail it is beneficial to first introduce some notation. At the *denoising* stage we assume that we observe a signal $x[n]$ which

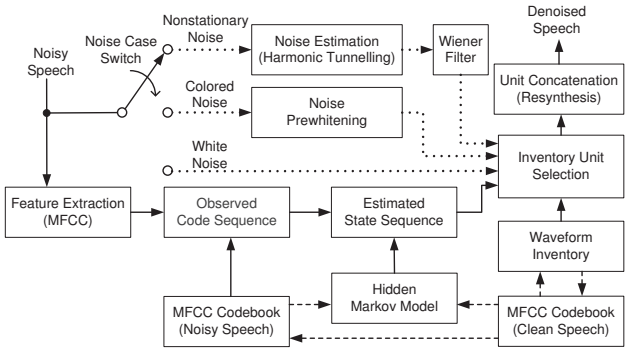

Figure 1. A block diagram of the proposed denoising method. Dashed lines indicate processing steps that are performed during system training. Dotted lines indicate signal preprocessing steps.

consists of speech $s[n]$ that is uttered by the enrolled speaker and is distorted by zero mean additive noise $v[n]$, i.e. $x[n] = s[n] + v[n]$.

At the *training* stage we use $\hat{s}[n]$ to, similarly, denote the *speaker enrollment data.* System training is done off-line from speaker-specific pre-recorded *clean* training signals. For simplicity we assume that all training records of speech are concatenated into one long training sequence $\hat{s}[n]$.

Throughout the paper we make use of speech *units* or *frames.* We represent a unit as a vector of N successive samples of a signal:

$$\mathbf{s}_n = [\ s[n-L]\ \ s[n-L+1]\ \ \ldots\ \ s[n-L+N-1]\ ]^{\mathrm{T}}. \quad (1)$$

The amount of overlap between adjacent frames is controlled by a step size L. If i denotes a unit (or frame) index then the associated vector is written as $\mathbf{s}_{iL}$. Symbols $\mathbf{x}_n$, $\mathbf{v}_n$, and $\hat{\mathbf{s}}_n$ are defined analogously to equation (1).

We use $\mathbb{S}$ to denote our *speech-waveform-unit inventory.* Set $\mathbb{S}$ consists of all clean training data frames $\hat{\mathbf{s}}_n$ ($\forall n$, i.e. with a step size of one) with the exception of data frames that are entirely silent[1].

Denoising is performed (up to an appropriate scaling factor) by finding a mapping $\mathbf{x}_{iL} \rightarrow \hat{\mathbf{s}}_{n(i)}$ that associates a specific inventory frame $\hat{\mathbf{s}}_{n(i)}$ to every observed noisy frame $\mathbf{x}_{iL}$. Note that this mapping is generally not fixed, but time-variant and context dependent. The resulting denoised signal $\tilde{s}[n]$ is obtained by "concatenating" the found frames $\hat{\mathbf{s}}_{n(i)}$ via a *sinusoidal model* based resynthesis technique [9]. The details of the procedure are described in section 2.2.

2.1. System Training and State Sequence Estimations

The goal of the system training stage is to provide the denoising procedure with an *inventory* of available speech units and a *hidden Markov model* that describes transition statistics within the inventory. During *inventory design* all inventory elements $\hat{\mathbf{s}}_n$ that belong to a similar *phonemic function*[2] are grouped into the same class. The purpose of the grouping is to be able to study the statistical properties of the group as a whole and then apply a resulting statistical description in the denoising process.

The details of the clustering methods that were used for the inventory design have to be omitted here due to space limitations. A comprehensive description can be found in a paper on our earlier work [4]. The only difference between the clustering method described in [4] and the one used in this work is that we enforced a strict separation between clusters containing voiced sections and clusters containing unvoiced sections. To maintain compatibility with the notation introduced in [4] we will refer to the resulting cluster sets of inventory vectors $\hat{\mathbf{s}}_n$ with $\mathbb{K}_k$ for $k = 1, 2, \ldots M$.

With the help of the inventory and its statistical description it becomes possible to define a sequence $k_{opt}(i)$ of "optimal" cluster memberships for incoming testing frames $\mathbf{x}_{iL}$. The sequence states that the "most likely" inventory element $\hat{\mathbf{s}}_{n(i)}$ to represent the denoised frame for $\mathbf{x}_{iL}$ is found in set $\mathbb{K}_{k_{opt}(i)}$. Again, the details of how to find the sequences $k_{opt}(i)$ have to be omitted. A comprehensive description is provided in [4].

2.2. Speech Denoising

The first step in implementing the actual denoising portion of the proposed method is the computation of the "optimal" cluster membership sequence $k_{opt}(i)$ as discussed in the previous section. After that, the two remaining tasks are: (1) the identification of the best match for each $\mathbf{x}_{iL}$ in $\mathbb{K}_{k_{opt}(i)}$, i.e. the *intra cluster frame matching,* and (2) the "concatenation" of the resulting inventory frames to resynthesize the targeted denoised signal.

2.2.1. Intra Cluster Frame Matching

We begin by defining a similarity measure between a noisy frame $\mathbf{x}_{iL}$ and an inventory element $\hat{\mathbf{s}}_n$. Deciding the best matched inventory element $\hat{\mathbf{s}}_n$ for the clean frame $\mathbf{s}_{iL}$ in noisy speech $\mathbf{x}_{iL}$ is essentially a signal detection problem. With a maximum likelihood criterion, if the additive noise $\mathbf{v}_{iL}$ is independent white Gaussian noise then a correlation detector should be used [10]. Since the power of the training frame and the testing frame may be significantly different, a power normalization is applied. We use $V^2 = \mathrm{E}\{\mathbf{v}_n^{\mathrm{T}}\mathbf{v}_n\}$ to denote the variance of the noise, and the estimated power $\sqrt{\|\mathbf{x}_{iL}\|^2 - V^2}$ of the underlying clean speech $s[n]$ is taken into account. For the estimate of the power of $s[n]$ we assume that the noise $v[n]$ is (approximately) orthogonal to $s[n]$. The similarity measure in the white Gaussian noise case is defined as follows:

$$\sigma(\mathbf{x}_{iL}, \hat{\mathbf{s}}_n) = \frac{\mathbf{x}_{iL}^{\mathrm{T}}\,\hat{\mathbf{s}}_n}{\sqrt{\|\mathbf{x}_{iL}\|^2 - V^2}\cdot\|\hat{\mathbf{s}}_n\|}. \quad (2)$$

If $\mathbf{v}_{iL}$ is colored noise then a prewhitening filter is used before the correlation detector. We assume that the linear prediction coefficients $a_1,\ a_2,\ \ldots,\ a_p$ of the noise $v[n]$ can be estimated via the *autocorrelation method.* The impulse response of the prewhitening filter $\mathbf{h}_w$ is then given by $\mathbf{h}_w = [\ 1\ a_1\ a_2\ \ldots\ a_p\ ]^T$. The similarity measure becomes:

$$\sigma(\mathbf{x}_{iL}, \hat{\mathbf{s}}_n) = \frac{(\mathbf{x}_{iL} * \mathbf{h}_w)^{\mathrm{T}}\,(\hat{\mathbf{s}}_n * \mathbf{h}_w)}{\sqrt{\|\mathbf{x}_{iL} * \mathbf{h}_w\|^2 - V_w^2}\cdot\|\hat{\mathbf{s}}_n * \mathbf{h}_w\|}, \quad (3)$$

where we use $V_w^2 = \mathrm{E}\{(\mathbf{v}_n * \mathbf{h}_w)^{\mathrm{T}}(\mathbf{v}_n * \mathbf{h}_w)\}$ to denote the variance of the prewhitened noise.

If $v[n]$ is non-stationary noise then a noise reduction stage for the voiced parts of the incoming speech signal is employed. We use *harmonic tunnelling* [11] to track the noise spectrum. The harmonic peaks of every incoming signal frame are detected using the approach described in [11]. The estimated noise spectrum is obtained in three steps. Firstly, we sample the spectrum in the tunnels between the harmonic spectral peaks. Secondly, we impose a weight function on the sampled spectrum to reduce the spectral smearing of

[1] We consider frames to be entirely silent if the total frame energy falls below a certain minimal level.

[2] We are using the term *phonemic function* in reference to a general, function carrying unit of a language. The group *may* or *may not* match with an actual *phoneme* defined for that language.

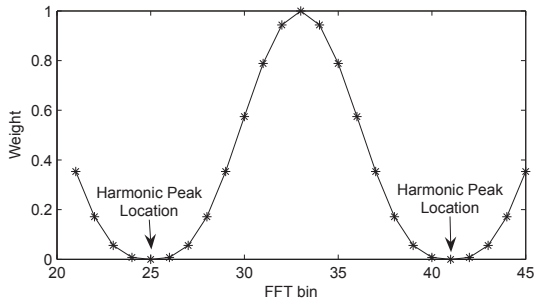

Figure 2. An illustration of the weight function employed in the *harmonic tunelling* approach.

harmonic peaks. The employed weight function is a sin^3 function as illustrated in figure 2. Finally the noise spectrum is smoothed in both time and frequency. The resulting estimated non-stationary noise spectrum was used in a Wiener filter to enhance the voiced part of the incoming speech sections. We use $\tilde{\mathbf{x}}_{iL}$ to denote the resulting enhanced speech frames. The similarity measure is then defined through equation (2) as $\sigma(\tilde{\mathbf{x}}_{iL}, \hat{\mathbf{s}}_n)$.

After the generation of an appropriate similarity measure between an incoming noisy frame $\mathbf{x}_{iL}$ (or $\tilde{\mathbf{x}}_{iL}$) and an inventory element $\mathbf{s}_n$ we can define an optimal intra cluster match $\hat{\mathbf{s}}^{(i,k)}$ via

$$\hat{\mathbf{s}}^{(i,k)} = \underset{\hat{\mathbf{s}}_n \in \mathbb{K}_k}{\arg\max}\ \sigma(\mathbf{x}_{iL}, \hat{\mathbf{s}}_n). \quad (4)$$

Denoising is simply accomplished by replacing each frame $\mathbf{x}_{iL}$ with the power normalized inventory frame $\alpha \cdot \hat{\mathbf{s}}^{(i,k_{opt}(i))}$:

$$\mathbf{x}_{iL} \quad \rightarrow \quad \alpha \cdot \hat{\mathbf{s}}^{(i,k_{opt}(i))}, \quad (5)$$

where the normalization factor $\alpha = \frac{\sqrt{\|\mathbf{x}_i\|^2 - V^2}}{\|\hat{\mathbf{s}}^{(i,k_{opt}(i))}\|}$.

2.2.2. Target Resynthesis

Lastly we concatenate the resulting inventory frames $\hat{\mathbf{s}}^{(i,k_{opt}(i))}$ to resynthesize the targeted denoised signal via a sinusoidal model expansion [9]. Reconcatenation with the sinusoidal model helps to minimize phase incompatibilities at the frame boundaries.

We extract the parameters of the peaks in the discrete Fourier transform (DFT) of frame $\hat{\mathbf{s}}^{(i,k_{opt}(i))}$ as amplitude $\hat{A}_l^i$, phase $\hat{\theta}_l^i$, and frequency $\hat{\omega}_l^i$, for $l = 1, 2, \ldots, Q(i)$. $Q(i)$ denotes the number of peaks below a threshold of 80. $\hat{A}_l^i$, $\hat{\theta}_l^i$, and $\hat{\omega}_l^i$ are referenced with respect to the *center* of each synthesis frame. We use the frequency-matching algorithm described in [9] to associate all of the parameters measured for an arbitrary frame i with a corresponding set of matching parameters for frame $i+1$. If we use $(\hat{A}_l^i, \hat{\theta}_l^i, \hat{\omega}_l^i)$ and $(\hat{A}_l^{i+1}, \hat{\theta}_l^{i+1}, \hat{\omega}_l^{i+1})$ to denote the parameters for the l^{th} frequency track in two consecutive frames, then we can interpolate the amplitude for the samples between the centers of those two frames via:

$$\hat{A}_l^i(k) = \hat{A}_l^i + \frac{\hat{A}_l^{i+1} - \hat{A}_l^i}{L} \cdot k \quad \text{for} \quad k = 0, 1, \cdots, L-1. \quad (6)$$

As described in [9], a cubic polynomial function can be used to interpolate the phase values $\hat{\theta}_l^i(k)$. After estimating and interpolating the above parameters for every sample of the incoming frame we can synthesize a signal $\tilde{s}[m]$ over the respective overlapping region of L samples with:

$$\tilde{s}[m] = \sum_{l=1}^{Q(i)} \hat{A}_l^i(m - iL) \cos[\hat{\theta}_l^i(m - iL)], \quad (7)$$

for $\quad iL \leq m \leq (i+1)L - 1$. The resulting $\tilde{s}[m]$ represents our targeted denoised speech signal.

3. EXPERIMENTAL RESULTS

The performance of the proposed method was evaluated with experiments over a subset of the `CMU_ARCTIC` database from the Language Technologies Institute at Carnegie Mellon University[3]. The `CMU_ARCTIC` database is specifically designed to be used with inventory based speech synthesis systems. The two subsets of the corpus employed in our study were the *US English* male speaker with identifier `BDL` and the *US English* female speaker with identifier `SLT`. The two subsets contain 1132 phonetically balanced English utterances each. Most utterances are between one and four seconds long. The data was appropriately low-pass filtered and subsampled to a processing sampling rate of 8 kHz. Additive noise was taken from the `NOISEX` database from the Institute for Perception-TNO, The Netherlands Speech Research Unit, RSRE, UK[4]. For our experiments we used *white* noise, *buccaneer jet cockpit* noise which is considered as colored noise, and *speech babble* noise which is considered as non-stationary noise at a signal to noise ratio of 10 dB.

From the two data sets we randomly chose 10 utterances each for testing and left the remaining 1122 utterances each for training. The training and testing sets were, thus, mutually disjoint. For the signal segmentation and inventory generation described in section 2 we used a frame length N of 160 samples (equivalent to 20 msec frames) and a step size L of 80 samples (equivalent to a 50% frame overlap). The size M of the employed inventory was 50 clusters. We used 40 clusters to model voiced frames and 10 clusters to model unvoiced frames.

An objective quality assessment was performed with the *Perceptual Evaluation of Speech Quality* (PESQ), the *Log Likelihood Ratio* (LLR), the *Itakura-Saito Distortion* (IS), and a *Cepstral Distance Measure* (CEPD). The PESQ measure, an ITU recommendation correlates very well with *subjective quality* of speech. Note that the LLR, the IS, and the CEPD are distortion measures (i.e. smaller values are better) whereas the PESQ is a quality measure (i.e. a bigger value is better). All measures are comprehensively described in the text by Loizou [12]. The quality/distortion measures evaluate the quality of the noisy or enhanced speech signal, using the original clean speech signal as the reference signal.

For benchmark purposes we computed results not only for the proposed method (**PM**) but also for four other standard and state-of-the-art methods. These methods are abbreviated with the two letter codes WF, MB, EM, and CB in tables I – III. Abbreviation **WF** denotes the *iterative Wiener filtering* scheme described in [1] (with 2 iterations). The code **MB** denotes the *multiband spectral subtraction* method proposed by Kamath and Loizou in 2002 [1]. **EM** represents the *minimum mean-square error log-spectral amplitude* estimator by Ephraim and Malah [13]. **CB** stands for a state-of-the-art *codebook-driven Wiener filtering* scheme similar to the one described in [7]. Note that the CB approach also required substantial training with speech and noise[5].

The experimental results are listed in tables I to III. The distortion/quality measures of the two best performing algorithms in each category are shown in a boldface font. It is readily visible that for the *white noise* and the *jet cockpit noise* our proposed method (PM) outperforms all other methods in all considered quality/distortion measures, especially in the PESQ measure (which is most significant in assessing *perceptual quality*).

[3] The corpus is available at <http://www.festvox.org/cmu_arctic>.

[4] The noise is available at <http://spib.rice.edu/spib/select_noise.html>.

[5] We used the same codebook clustering method for the proposed method and the **CB** method to make the comparison fair between the two approaches.

Table I
Average Objective Quality Measures for Conventional Enhancement Methods and the Proposed Method under Additive White Noise at 10dB SNR.

	Quality Measures	Noisy	PM	WF	MB	EM	CB
B	PESQ	1.76	**2.70**	**2.53**	2.23	2.51	**2.53**
D	LLR	1.02	**0.61**	1.02	0.88	**0.64**	1.03
L	IS	1.50	**1.24**	2.99	2.24	1.93	**1.78**
	CEPD	5.55	**4.21**	5.81	5.09	**4.31**	5.68
S	PESQ	1.69	**2.75**	2.32	2.24	**2.54**	2.36
L	LLR	1.29	**0.47**	0.80	0.78	**0.65**	0.84
T	IS	2.02	**1.24**	3.98	1.44	1.65	**1.33**
	CEPD	6.74	**4.09**	5.30	4.94	**4.58**	5.06

Table II
Average Objective Quality Measures for Conventional Enhancement Methods and the Proposed Method under Additive Jet Cockpit Noise at 10dB SNR.

	Quality Measures	Noisy	PM	WF	MB	EM	CB
B	PESQ	2.01	**2.73**	2.47	2.44	2.58	**2.62**
D	LLR	0.74	**0.51**	0.90	0.72	**0.54**	0.90
L	IS	1.14	**0.84**	7.58	2.10	1.74	**1.72**
	CEPD	4.63	**3.84**	5.13	4.50	**3.96**	4.84
S	PESQ	1.88	**2.75**	2.15	2.37	**2.62**	2.36
L	LLR	0.97	**0.58**	0.83	0.73	**0.62**	0.86
T	IS	1.59	**0.96**	10.68	**1.50**	1.86	1.87
	CEPD	5.95	**4.04**	5.45	4.80	**4.66**	4.77

Table III
Average Objective Quality Measures for Conventional Enhancement Methods and the Proposed Method under Additive Babble Noise at 10dB SNR.

	Quality Measures	Noisy	PM	WF	MB	EM	CB
B	PESQ	2.32	**2.60**	2.15	**2.66**	2.51	2.40
D	LLR	0.74	0.61	0.83	**0.35**	**0.38**	0.65
L	IS	1.14	1.46	10.78	**0.60**	**0.65**	7.50
	CEPD	4.63	4.21	**2.45**	2.87	2.98	**2.52**
S	PESQ	2.14	**2.64**	2.24	**2.49**	2.46	2.35
L	LLR	0.49	0.51	0.52	**0.48**	**0.47**	0.48
T	IS	0.74	1.64	3.50	**1.09**	**1.00**	1.39
	CEPD	3.80	4.32	3.90	**3.71**	**3.72**	3.85

The proposed method still achieves top performances in PESQ measure for non-stationary *speech babble noise.* Only in the male speaker case (BDL) was the proposed method slightly inferior in PESQ to the *multiband spectral estimation method* (MB). The performance of the proposed method was somewhat less successful in terms of the other distortion measures, especially in comparison to its main competitor the MB case. Informal listening tests with a small group of subjects, however, revealed that the true perceptual quality of the proposed method was at least comparable (if not slightly superior) to the MB method in this case.

The improvements of the proposed method come at the cost of an increased complexity. The complexity of the proposed method is dominated by the intra-cluster search and grows at an order of $K \log K$ with K being the sample-number per cluster (in average $K = 384000$ in our experiments).

4. CONCLUSIONS

We presented a new method for the denoising of speech. Our approach is based on an *inventory style* speech re-synthesis scheme that utilizes a statistical analysis of the underlying parameter space. The required statistical descriptions were obtained from noise enrollment and from speaker enrollment in clean conditions. With experiments we have shown that the proposed method performs very well in comparison to commonly used waveform based denoising methods.

5. REFERENCES

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

[2] Y. Hu and P. C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 1, pp. 59–67, Jan. 2004.

[3] A. Mouchtaris, J. Van der Spiegel, P. Mueller, and P. Tsakalides, "A spectral conversion approach to single-channel speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1280–1193, May 2007.

[4] X. Xiao, P. Lee, and R. M. Nickel, "Inventory based speech denoising with hidden Markov models," in *EUSIPCO 2008*.

[5] E. Zavarehei, S. Vaseghi, and Q. Yan, "Noisy speech enhancement using harmonic-noise model and codebook-based post-processing," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1194–1203, May 2007.

[6] D. Y. Zhao and W. B. Kleijn, "HMM-based gain modeling for enhancement of speech in noise," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 882–892, 2007.

[7] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 163–176, Jan. 2006.

[8] D. O'Shaughnessy, "Modern methods of speech synthesis," *IEEE Circuits and Systems Magazine*, vol. 7, no. 3, pp. 6–23, 2007.

[9] T. F. Quatieri, *Discrete-Time Speech Signal Processing*, Prentice Hall, Upper Saddle River, NJ 07458, 2002.

[10] H.Vincent Poor, *An Introduction to Signal Detection and Estimation*, Springer-Verlag, 1994.

[11] D. Ealey, H. Kelleher, and D. Pearce, "Harmonic tunnelling: tracking non-stationary noises during speech," *EUROSPEECH*, pp. 437–440, Sept. 2001.

[12] Y. Hu and P. Loizou, "Evaluation of objective measures for speech enhancement," *Proceedings of INTERSPEECH-2006*, Sept. 2006.

[13] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. 33, pp. 443–445, Apr. 1985.