A SPECTRAL CONVERSION APPROACH TO THE ITERATIVE WIENER FILTER FOR SPEECH ENHANCEMENT

*Athanasios Mouchtaris, Jan Van der Spiegel**

Department of Electrical and Systems Engineering
University of Pennsylvania
Philadelphia, PA 19104

Paul Mueller

Corticon Inc.
155 Hughes Rd.
King of Prussia, PA 19406

ABSTRACT

The Iterative Wiener Filter (IWF) for speech enhancement in additive noise is an effective and simple algorithm to implement. One of its main disadvantages is the lack of proper criteria for convergence, which has been shown to introduce severe degradation to the estimated clean signal. Here, an improvement of the IWF algorithm is proposed, when additional information is available for the signal to be enhanced. If a small amount of clean speech data is available, spectral conversion techniques can be applied for estimating the clean short-term spectral envelope of the speech signal from the noisy signal, with significant noise reduction. Our results show an average improvement compared to the original IWF that can reach 2 dB in the segmental output Signal-to-Noise Ratio (SNR), in low input SNR's, which is perceptually significant.

1. INTRODUCTION

According to the Iterative Wiener Filter (IWF) algorithm [1] for speech enhancement in additive noise, the non-causal Wiener filter [2] is applied to the noisy speech iteratively, while the spectral estimate of the speech signal is based on all-pole modeling of the enhanced signal at each iteration. One of its main disadvantages is the lack of proper criteria for convergence, which has been shown to introduce severe degradation to the estimated clean signal, an issue that has been mainly attacked by introducing constraints during the all-pole estimation [3, 4]. In this paper, we show that spectral conversion techniques can be applied to the speech enhancement problem within the IWF framework, with the assumption that a small amount of clean speech data is initially available. Spectral conversion has been applied previously to voice conversion, whose objective is to modify the speech characteristics of a particular speaker in such manner, as to sound like speech by a different target speaker [5, 6, 7]. Here, we show that there are valid analogies between these two different fields of speech processing, which can be exploited and produce satisfactory results. More specifically, we view the problem of speech enhancement in additive noise as similar to voice conversion, where the source speech is the noisy speech, and the target speech is the clean speech. There are some difficulties that arise from this assumption, that are detailed next.

The common characteristic of voice conversion approaches is that they focus on the short-term spectral properties of the speech signals, which they modify according to a conversion function designed during the training phase. This was in fact our motivation for applying these approaches to the IWF, whose performance greatly depends on the estimation of the short-term properties of the speech signal. During training, the parameters of this conversion function are derived based on minimizing some error measure. In order to achieve this, a speech corpus is needed that contains the same utterances (words, sentences, *etc.*) from both the source and target speakers (parallel corpus). If a corpus containing the same utterances of the noisy and clean speech is available, we show that spectral conversion applies to this problem favorably.

In Fig. 1, the block diagram of the proposed algorithms is given. During the first iteration of the IWF algorithm, the all-pole coefficients of the speech are directly derived from the noisy signal. Spectral conversion is then applied, resulting in better estimation of the clean speech parameters based on the training phase. After the first iteration, the IWF algorithm proceeds as usual, although our simulations showed that additional iterations do not offer significant improvement in most cases. For parallel training, clean and noisy speech data are required, with the additional constraint that the same utterances must be available from the clean and noisy speech. It is often difficult or even impossible to collect such a corpus. In [8], we proposed a conversion algorithm that relaxes this constraint. Our approach was to adapt the conversion parameters for a given pair of source and target speakers, to the particular pair of speakers for which no parallel corpus is available. Similarly here, we assume that a parallel corpus is available for noisy speech 2 and clean speech 2 in Fig. 1, and for this pair a conversion function is derived by employing a conversion method given in the literature [7]. For the particular pair of clean and noisy speech that we focus on, a non-parallel corpus is available for training. Constrained adaptation techniques allow for deriving the needed conversion parameters by relating the non-parallel corpus to the parallel corpus. We show that the speaker and noise characteristics in the two pairs of speech data can differ, while noise stationarity is assumed. The training phase is greatly simplified with this latter approach, since only few sentences of clean speech are needed, while the noisy speech is readily available.

2. ITERATIVE WIENER FILTER

For the case examined here, the noisy signal $y(n)$ is given by

$$y(n) = s(n) + d(n) \tag{1}$$

where $s(n)$ is the clean speech signal and $d(n)$ is the uncorrelated with $s(n)$ additive noise. The IWF algorithm estimates the speech signal from noisy speech by iteratively applying the non-causal Wiener filter

$$H(\omega) = \frac{P_s(\omega)}{P_s(\omega) + P_d(\omega)} \tag{2}$$

where $H(\omega)$ denotes the frequency response of the filter, $P_s(\omega)$ is the power spectral density (psd) of $s(n)$ and $P_d(\omega)$ is the psd of $d(n)$. The psd of the speech signal in IWF is estimated from

$$P_s(\omega) = \frac{G^2}{\left|1 + \sum_{m=1}^{p} a(m) e^{-j\omega m}\right|^2} \tag{3}$$

*This research has been funded by the Catalyst Foundation.

0-7803-8603-5/04/$20.00 ©2004 IEEE

i.e. the all-pole model of order p of the noisy speech, while the psd of the noise can be estimated from the noisy speech during regions of silence. The constant term G can be estimated from the energy difference between the noisy signal and the estimated noise. The algorithm operates in short-time segments of the speech signal, and a new filter applies for each segment. Usually a small number of iterations for each segment is required for convergence, so the computational requirements of the algorithm are limited. However, there is no proper criterion for convergence of the IWF procedure, which is an important disadvantage since it has been shown that after a few iterations the solution greatly deviates from the correct estimate. Towards addressing this issue, several improvements have been proposed that constrain the all-pole estimate at each iteration so that the parameters retain speech-like properties.

3. SPECTRAL CONVERSION

From the reference and target training waveforms, we extract the parameters that model their short-term spectral properties (in this paper we use the line spectral frequencies - LSF's - due to their desirable interpolation properties [7]). This results in two vector sequences, $[\boldsymbol{x}_1\boldsymbol{x}_2\ldots\boldsymbol{x}_n]$ and $[\boldsymbol{y}_1\boldsymbol{y}_2\ldots\boldsymbol{y}_n]$, of reference and target spectral vectors respectively. The objective of spectral conversion methods is to derive a function $\mathcal{F}(\cdot)$ which, when applied to vector $\boldsymbol{x}_k$, produces a vector close in some sense to vector $\boldsymbol{y}_k$. For the noise enhancement problem, the vector sequence $\boldsymbol{x}_k$ corresponds to the noisy speech, while the sequence $\boldsymbol{y}_k$ corresponds to the clean speech. Gaussian mixture models (GMM's) have been successfully applied to the voice conversion problem [6, 7]. GMM's approximate the unknown probability density function (pdf) of a random vector $\boldsymbol{x}$ as a mixture of Gaussians whose parameters (mean vectors, covariance matrices and prior probabilities of each Gaussian class), can be estimated from the observed data using the expectation maximization (EM) algorithm [9].

We focus on the spectral conversion method of [7], which offers great insight as to what the conversion parameters represent. Assuming that $\boldsymbol{x}$ and $\boldsymbol{y}$ are jointly Gaussian for each class ω_i, then, in mean-squared sense, the optimal choice for the function $\mathcal{F}$ is

$$\begin{aligned}\mathcal{F}(\boldsymbol{x}_k) &= \mathrm{E}(\boldsymbol{y}\,|\,\boldsymbol{x}_k) \qquad (4)\\ &= \sum_{i=1}^{M} p(\omega_i\,|\,\boldsymbol{x}_k)\left[\boldsymbol{\mu}_i^{y} + \boldsymbol{\Sigma}_i^{yx}\boldsymbol{\Sigma}_i^{xx\,-1}\left(\boldsymbol{x}_k - \boldsymbol{\mu}_i^{x}\right)\right],\end{aligned}$$

where $\mathrm{E}(\cdot)$ denotes the expectation operator and the conditional probabilities $p(\omega_i\,|\,\boldsymbol{x}_k)$ are given from

$$p(\omega_i\,|\,\boldsymbol{x}_k) = \frac{p(\omega_i)\mathcal{N}(\boldsymbol{x}_k;\boldsymbol{\mu}_i^{x},\boldsymbol{\Sigma}_i^{xx})}{\sum_{j=1}^{M} p(\omega_j)\mathcal{N}(\boldsymbol{x}_k;\boldsymbol{\mu}_j^{x},\boldsymbol{\Sigma}_j^{xx})}. \qquad (5)$$

All the parameters in the two above equations are estimated using the EM algorithm on the joint model of $\boldsymbol{x}$ and $\boldsymbol{y}$. In practice this means that the EM algorithm is performed during training on the concatenated vectors $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$. A time-alignment procedure is required in this case, and this is only possible when a parallel corpus is used. Another issue is that performance considerations, when using the adaptation procedure described in the next paragraph, dictate that the covariance matrices used in this conversion method be of diagonal form. In order to achieve this restriction some issues must be addressed due to the joint model used [10].

4. CONSTRAINED GMM ESTIMATION

In the previous section we described the spectral conversion algorithm that can result in estimates of the clean speech spectral features from the noisy speech. These estimates can then be directly used in the IWF algorithm when applied to (3) during the first iteration. However, a parallel training corpus is needed for this method, which sometimes might be difficult to acquire. As an alternative, we propose in this section a procedure which is based on the spectral conversion method of the previous paragraph, but allows for a non-parallel corpus. We show that this is possible under the assumption that a parallel speech corpus is available for a *different* reference and target speech pair (*i.e.* different speaker and noise conditions). In order to achieve this result, we apply the maximum-likelihood constrained adaptation method [11], which offers the advantage of a simple probabilistic linear transformation leading to a mathematically tractable solution.

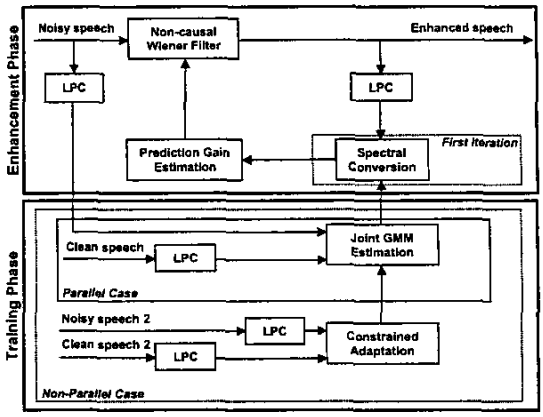

Fig. 1. Block diagram outlining spectral conversion for a parallel and non-parallel corpus within the IWF framework. Non-parallel training is achieved by adaptation of the parameters derived from parallel training of a different speaker and noise conditions.

As mentioned, we also assume that a parallel speech corpus is available for a different speaker and noise conditions, in addition to the particular pair of speaker and noise for which only a non-parallel corpus exists. From the parallel corpus, we obtain a joint GMM model, derived as explained in Section 3. The spectral vectors that correspond to the reference speech are considered as realizations of random vector $\boldsymbol{x}$, while $\boldsymbol{y}$ corresponds to the target speech of the parallel corpus. From the non-parallel corpus, we also obtain a sequence of spectral vectors, considered as realizations of random vector $\boldsymbol{x}'$ for the reference speech and $\boldsymbol{y}'$ for the target speech. We then relate the random variables $\boldsymbol{x}'$ and $\boldsymbol{x}$, as well as $\boldsymbol{y}'$ and $\boldsymbol{y}$, in order to derive a conversion function for the non-parallel corpus based on the parallel corpus parameters.

We assume that the target random vector $\boldsymbol{x}'$ is related to reference random vector $\boldsymbol{x}$ by a probabilistic linear transformation

$$\boldsymbol{x}' = \mathbf{A}_j\boldsymbol{x} + \boldsymbol{b}_j \;\text{ with probability } p(\lambda_j\;\omega_i),\quad j = 1,\ldots,N. \qquad (6)$$

Each of the component transformations j is related with a specific Gaussian i of $\boldsymbol{x}$ with probability $p(\lambda_j\;\omega_i)$ satisfying

$$\sum_{j=1}^{N} p(\lambda_j\;\omega_i) = 1,\quad i = 1,\ldots,M. \qquad (7)$$

In the above equations M is the number of Gaussians of the GMM that corresponds to the joint vector sequence of the parallel corpus, $\mathbf{A}_j$ is a $K \times K$ matrix (K is the dimensionality of $\boldsymbol{x}$), and $\boldsymbol{b}_j$ is a vector of the same dimension with $\boldsymbol{x}$. Random vectors $\boldsymbol{y}'$ and $\boldsymbol{y}$ are related by another probabilistic linear transformation, similar to (6), where matrix $\mathbf{A}_j$ is now substituted by $\mathbf{C}_\rho$, vector $\boldsymbol{b}_j$ becomes $\boldsymbol{d}_\rho$, and $p(\lambda_j\;\omega_i)$ becomes $p(\kappa_\rho\;\omega_i)$. Note that classes ω_i

Method	IWF	Ideal	SS	SC	SC-A
ASSNR	5.6839	7.3654	3.0678	6.8538	6.7877

Table 1. Resulting ASSNR (dB) for input SNR of 0 dB for Iterative Wiener Filter (IWF), perfect prediction (Ideal), Spectral Subtraction (SS), Spectral Conversion with IWF (SC), and Spectral Conversion followed by adaptation and IWF (SC-A).

are the same for $\boldsymbol{x}$ and $\boldsymbol{y}$ by design in Section 3. All the unknown parameters can be estimated by use of the non-parallel corpus and the GMM of the parallel corpus, by applying the EM algorithm. Based on the linearity of the transformations and the fact that for a specific class the pdf's are Gaussian, it can be shown [8], that the conversion function for the non-parallel case is

$$\begin{aligned} \mathcal{F}(\boldsymbol{x}_k') &= \mathrm{E}(\boldsymbol{y}' \,\boldsymbol{x}_k') \qquad (8) \\ &= \sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{\rho=1}^{L} p(\omega_i \,\boldsymbol{x}_k')p(\lambda_j \,\boldsymbol{x}_k', \omega_i)p(\kappa_\rho \,\omega_i) \\ &\quad \left[\mathbf{C}_\rho\boldsymbol{\mu}_i^y + \boldsymbol{d}_\rho + \mathbf{C}_\rho\boldsymbol{\Sigma}_i^{yx}{\boldsymbol{\Sigma}_i^{xx}}^{-1}\mathbf{A}_j^{-1}\right. \\ &\quad \left.\left(\boldsymbol{x}_k' - \mathbf{A}_j\boldsymbol{\mu}_i^x - \boldsymbol{b}_j\right)\right], \end{aligned}$$

$$p(\omega_i \,\boldsymbol{x}_k') = \frac{p(\omega_i)\sum_{j=1}^{N} p(\lambda_j \,\omega_i)\mathrm{g}(\boldsymbol{x}_k' \,\omega_i, \lambda_j)}{\sum_{i=1}^{M}\sum_{j=1}^{N} p(\omega_i)p(\lambda_j \,\omega_i)\mathrm{g}(\boldsymbol{x}_k' \,\omega_i, \lambda_j)}, \qquad (9)$$

$$p(\lambda_j \,\boldsymbol{x}_k', \omega_i) = \frac{p(\lambda_j \,\omega_i)\mathrm{g}(\boldsymbol{x}_k' \,\omega_i, \lambda_j)}{\sum_{j=1}^{N} p(\lambda_j \,\omega_i)\mathrm{g}(\boldsymbol{x}_k' \,\omega_i, \lambda_j)}, \qquad (10)$$

$$\mathrm{g}(\boldsymbol{x}' \,\omega_i, \lambda_j) = \mathcal{N}(\boldsymbol{x}'; \mathbf{A}_j\boldsymbol{\mu}_i^x + \boldsymbol{b}_j, \mathbf{A}_j\boldsymbol{\Sigma}_i^{xx}\mathbf{A}_j^T). \qquad (11)$$

5. SIMULATION RESULTS

In this section we test the performance of the spectral conversion and adaptation methods described in the previous paragraphs to the speech enhancement problem within the IWF framework. We use 40 ms. windows (the sampling rate is 22.050 kHz) and the spectral vectors used here are the LSF's (28^{th} order) due to their favorable interpolation properties. For these experiments we use white Gaussian noise, while preliminary results verify the validity of our methods to various types of noise, such as pink and car noise. The error measure employed is the output average segmental SNR,

$$\mathrm{ASSNR(dB)} = \frac{1}{n}\sum_{k=1}^{n} 10\log_{10}\left(\frac{\boldsymbol{x}_k^T\boldsymbol{x}_k}{(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k)^T(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k)}\right),$$

where $\boldsymbol{x}_k$ is the clean speech signal for segment k, and $\hat{\boldsymbol{x}}_k$ is the estimated speech signal for segment k. We test the performance of the algorithms using the ASSNR for various values of input (global) SNR. The corpus used is the *VOICES* corpus, available from OGI's CSLU [12, 13]. This is a parallel corpus and is used for both the parallel and non-parallel training cases that are examined in this section, in a manner explained in the next paragraph. We test the performance of the two algorithms proposed here (one case (4) for parallel training and one (8) for non-parallel training), in comparison to the unconstrained IWF and spectral subtraction [14]. The ideal error for the IWF method is given as well (*i.e.* perfect prediction of the all-pole coefficients, which are available only in the simulation environment). It is important to note that the corpus used contains a total of 50 sentences, of which a total of 40 is used for training purposes (as explained next) and the remaining 10 are used for testing. All the results given in this section are averaged over these 10 sentences and, in addition, for each sentence the result is the average of 10 different realizations of noise.

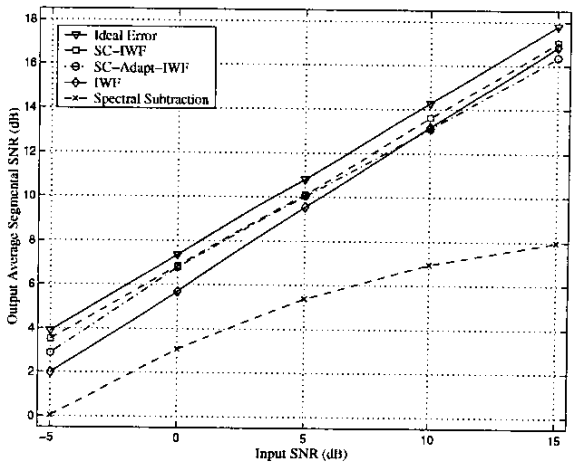

Fig. 2. Resulting ASSNR (dB) for different values of input SNR, for the five cases tested, *i.e.* perfect prediction (Ideal Error), the Iterative Wiener Filter (IWF), Spectral Conversion for IWF (SC-IWF, parallel corpus), Spectral Conversion by adaptation for IWF (SC-Adapt-IWF, non-parallel corpus), and Spectral Subtraction.

In Fig. 2, the ASSNR is given for the five cases tested, for various values of input SNR. As mentioned in the previous paragraph, we test the two algorithms proposed here (for parallel training (SC-IWF) and non-parallel training (SC-Adapt-IWF)), compared with the IWF algorithm, spectral subtraction, and the theoretically best possible performance of the IWF. For SC-IWF, the number of GMM parameters for training is 16 and the number of vectors in training is 5,000, which corresponds to about 15 sentences. For SC-Adapt-IWF, the number of adaptation parameters is 4 ($L = N = 4$), and the number of training vectors is 5,000. From the figure it is evident that the SC-IWF algorithm improves on the IWF algorithm, especially in low input SNR's, which is exactly what is desired. In many cases in our simulations the performance improvement reached 2 dB, which is quite significant perceptually in low SNR's. The SC-IWF algorithm can only be implemented when a parallel training dataset is available. When this is not possible, the SC-Adapt-IWF method was proposed, which is based on adapting the conversion parameters of a different pair of speaker/noise conditions. In this figure, we plot the performance of the SC-Adapt-IWF algorithm based on a different speaker from our corpus in white Gaussian noise of 10 dB SNR. We can conclude that the adaptation is very successful in low SNR's, when it performs only marginally worse than SC-IWF. In higher SNR's the training corpus, parallel or non-parallel, does not seem to offer any advantage when compared to IWF, which is sensible since the all-pole parameters can be estimated by the IWF quite efficiently in this low-noise case. The results for input SNR of 0 dB are also given in Table 1 for comparison with the results in Tables 2 and 3.

In Table 2, the ASSNR is given for the parallel case (SC-IWF) for 0 dB input SNR, for various numbers of GMM parameters and vectors in training. When comparing the performance of the various numbers of GMM parameters, the vectors in training are 5,000. We can see from the table that when increasing the number of GMM parameters in training, the performance of the algorithm improves as expected (since this corresponds to more accurate modeling of the spectral vectors). We must keep in mind that a 0.5 dB improvement is perceptible in low SNR. For the second case examined in this table, namely the effect of the training dataset size on the performance of the algorithm, the number of

GMM's	2	4	8	16	32
ASSNR	6.3655	6.4737	6.7932	6.8538	6.8966
Vectors	500	1000	2000	5000	10000
ASSNR	6.5838	6.7402	6.8172	6.8538	7.0362

Table 2. Resulting ASSNR in dB (parallel training, 0 dB input SNR), for different numbers of GMM parameters (for 5,000 vectors) and training vectors (for 16 GMM parameters).

GMM parameters is 16. From the table we can see that the performance of the algorithm improves when more training vectors are available, although not significantly for more than 2,000 vectors. The fact that only a small number of training data results in significant improvement over IWF is important, since this corresponds to requiring only a small amount of clean speech data.

In Table 3, the ASSNR is given for the non-parallel case and input SNR of 0 dB, for various choices of adaptation parameters (again, in (8) $L = N$) and training dataset size. When varying the number of adaptation parameters, the training dataset contains 5,000 vectors, and when varying the number of vectors in the training dataset, the number of adaptation parameters is $L = N = 4$. It is important to note that for all cases examined, the sentences used for adaptation are different than those used to obtain the conversion parameters (*i.e.* from the different speaker and noise conditions, for which a parallel corpus is used with 16 GMM parameters and 5,000 training vectors). From the table we can see that increasing the number of adaptation parameters improves the algorithm performance, which is an intuitive result since a larger number of adaptation parameters better models the statistics of the spectral vectors. Adaptation of 0 parameters corresponds to the case when no adaptation takes place, *i.e.* when the derived parameters for a different speaker and noise conditions are applied to the non-parallel case. It is evident that adaptation is indeed required, reducing the error considerably. Performance improvement is also noticed when increasing the number of training data, noting again that only few training data can produce desirable results.

It is important to note that the results given here correspond to the ideal case when it is known when the IWF algorithm converges. In reality, proper convergence criteria for the IWF algorithm do not exist, and as mentioned this can severely degrade its performance. In contrast, the spectral conversion based algorithms proposed here were found to not require additional iterations for achieving minimal error. This should be expected since the spectral conversion methods result in a good approximation of the all-pole parameters of the clean speech, thus no significant improvement is achieved with additional iterations. This is an important advantage of the proposed algorithms when compared to other IWF-based speech enhancement methods. Another issue is that in segments of very low speech energy, resulting in very low SNR, the methods proposed here might result in abrupt noise. These cases can be identified by applying a threshold, derived from the noisy speech energy as a pre-processing step.

6. CONCLUSIONS

In this paper we applied spectral conversion techniques, originally developed for voice conversion, to the speech enhancement problem. The two algorithms given here, one for parallel and one for non-parallel training, can estimate the clean speech all-pole parameters from the noisy signal. These parameters can then be applied within the IWF framework, instead of the IWF initial parameter estimation directly from the noisy signal. The results verified that this spectral conversion approach results in a better estimate of the clean speech, with the additional advantage that the iterative estimation procedure, a major drawback for the IWF algorithm, can usually be circumvented with minimal effects on the performance

Param.	0	1	2	4	6
ASSNR	6.2211	6.6679	6.7513	6.7877	6.6805
Vectors	500	1000	2000	5000	10000
ASSNR	5.9452	6.7106	6.7404	6.7877	6.8525

Table 3. Resulting ASSNR in dB (non-parallel training, 0 dB input SNR), for different numbers of adaptation parameters (for 5,000 vectors) and training vectors (for 4 adaptation parameters).

of the proposed algorithms. Our future plans include testing our methods for various types of stationary and quasi-stationary noise.

7. REFERENCES

[1] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. ASSP-26, pp. 197–210, June 1978.

[2] A. Papoulis, *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 1991.

[3] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 795–805, April 1991.

[4] T. V. Sreenivas and P. Kirnapure, "Codebook constrained Wiener filtering for speech enhancement," *IEEE Trans. Speech and Audio Processing*, vol. 4, pp. 383–389, September 1996 1996.

[5] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *IEEE Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, (New York, NY), pp. 655–658, April 1988.

[6] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, vol. 6, pp. 131–142, March 1998.

[7] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *IEEE Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, (Seattle, WA), pp. 285–289, May 1998.

[8] A. Mouchtaris, J. Van der Spiegel, and P. Mueller, "Non-parallel training for voice conversion by maximum likelihood constrained adaptation," To appear *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2004.

[9] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 72–83, January 1995.

[10] A. Mouchtaris, S. S. Narayanan, and C. Kyriakakis, "Maximum likelihood constrained adaptation for multichannel audio synthesis," in *Conf. Record of the Thirty-Sixth Asilomar Conf. Signals, Systems and Computers*, vol. 1, (Pacific Grove, CA), pp. 227–232, November 2002.

[11] V. D. Diakoloukas and V. V. Digalakis, "Maximum-likelihood stochastic-transformation adaptation of Hidden Markov Models," *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 177–187, March 1999.

[12] A. Kain, *High Resolution Voice Transformation*. PhD thesis, OGI School of Science and Engineering at Oregon Health and Science University, October 2001.

[13] http://www.cslu.ogi.edu/corpora/voices/.

[14] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. ASSP-27, pp. 113–120, April 1979.