A Spectral Conversion Approach to Single-Channel Speech Enhancement

Athanasios Mouchtaris, *Member, IEEE*, Jan Van der Spiegel, *Fellow, IEEE*, Paul Mueller, and Panagiotis Tsakalides, *Member, IEEE*

***Abstract*—In this paper, a novel method for single-channel speech enhancement is proposed, which is based on a spectral conversion feature denoising approach. Spectral conversion has been applied previously in the context of voice conversion, and has been shown to successfully transform spectral features with particular statistical properties into spectral features that best fit (with the constraint of a piecewise linear transformation) different target statistics. This spectral transformation is applied as an initialization step to two well-known single channel enhancement methods, namely the iterative Wiener filter (IWF) and a particular iterative implementation of the Kalman filter. In both cases, spectral conversion is shown here to provide a significant improvement as opposed to initializations using the spectral features directly from the noisy speech. In essence, the proposed approach allows for applying these two algorithms in a user-centric manner, when "clean" speech training data are available from a particular speaker. The extra step of spectral conversion is shown to offer significant advantages regarding output signal-to-noise ratio (SNR) improvement over the conventional initializations, which can reach 2 dB for the IWF and 6 dB for the Kalman filtering algorithm, for low input SNRs and for white and colored noise, respectively.**

***Index Terms*—Gaussian mixture model (GMM), parameter adaptation, spectral conversion, speech enhancement.**

I. Introduction

SPECTRAL conversion has the objective of estimating spectral parameters with specific target statistics from spectral parameters with specific source statistics, using training data as a means of deriving the estimation parameters. Spectral conversion has been defined within the voice conversion problem, where the objective is to modify the speech characteristics of a particular speaker in such manner, as to sound like speech by a different target speaker (for example [1]–[5] and references therein). In this paper, we have applied spectral conversion to the speech (in additive noise) enhancement problem, by considering this problem as analogous to voice conversion, where the source speech is the noisy speech, and the target speech is the clean speech, the noise being either white or colored, and possibly nonstationary. In essence, we practically demonstrate that spectral conversion can be viewed as a very useful estimation method outside the context of voice conversion. Our objective is to apply spectral conversion as a feature denoising method for speech enhancement, within a linear filtering framework (Wiener and Kalman filtering are examined). Although it is possible to directly use the converted features for synthesizing an enhanced speech signal (using the noisy speech residual), our observation has been that we can obtain perceptually better speech quality when we use the new features as a means for estimating the parameters of an "optimal" linear filter.

The single-channel speech enhancement problem has received wide attention and consequently numerous algorithms have been proposed on the subject. In this paragraph, we give a brief overview of the most influential research directions that have been proposed over the years. Concentrating on the additive noise problem, one of the most popular, effective, and simple algorithms to implement is spectral subtraction [6]. According to this method, the speech signal is processed in short-term segments, and the noise statistics are estimated from segments for which no speech is available. For the segments where speech is available, the estimated noise is subtracted in the frequency domain from the noisy signal. The method although simple is quite effective. However, a significant disadvantage is that some noise frequencies remain unaffected in the enhanced speech resulting in tonal noise (or musical noise). The iterative Wiener filter (IWF) method has been proposed [7], which also operates on short-term segments of the speech signal. The method estimates the clean speech all-pole parameters iteratively, and then applies an approximated noncausal Wiener filter [8] at each iteration; IWF has been shown to reduce the error after each iteration and asymptotically converge to the true noncausal Wiener filter. The disadvantage of this method is that no proper convergence criteria exist, and after just a few iterations beyond convergence, the quality of the speech estimate becomes degraded. Methods have been suggested that partly address this issue by introducing constraints to the estimated all-pole speech parameters, so that they retain speech-like properties [9], [10]. Other main directions on the problem include estimation theoretic approaches such as minimum mean-squared estimation of the optimal linear filter [including hidden Markov model (HMM)-based approaches] [11]–[14], subspace-based methods [15] where the enhancement is based on estimating the signal and noise subspaces

Manuscript received December 22, 2005; revised December 4, 2006. This work was supported in part by the General Secretariat for Research and Technology of Greece and the European Social Fund, Program ΕΠΑΝ Code 05NON-EU-1 and in part by a Marie Curie International Reintegration Grant within the Sixth European Community Framework Program. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hong-Goo Kang.

A. Mouchtaris and P. Tsakalides are with the Computer Science Department, University of Crete, Heraklion, Crete 71409, Greece, and also with the Institute of Computer Science, Foundation for Research and Technology-Hellas (FORTH-ICS), Heraklion, Crete 71110, Greece (e-mail: mouchtar@ieee.org; tsakalid@ics.forth.gr).

J. Van der Spiegel is with the Electrical and Systems Engineering Department, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: jan@seas.upenn.edu).

P. Mueller is with Corticon, Inc., King of Prussia, PA 19406 USA (e-mail: cortion@aol.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2007.894511

and subsequent estimation of the optimal in some sense filter, Kalman filtering approaches [16], [17], taking advantage of particular speech models, and perceptual-based enhancement methods, where the noise is suppressed by exploiting properties of the human auditory system [18].

Use of spectral conversion for speech enhancement produces better estimates of the speech spectral features at the expense of the requirement for training data. In many practical scenarios, however, it is possible to have *a priori* access to clean speech signals, and many popular algorithms for speech enhancement have been developed under this assumption, such as HMM-based algorithms [13], [14]. A significant similarity of such approaches with the methods presented in this paper, is the use of mixture models for the probability density function (pdf) of the spectral features. In contrast with many corpus-based approaches, our spectral conversion methods do not assume any model for the background noise and do not require any noise training data. Our methods (in addition to the clean speech signals) require access to the noisy speech signal for training, which is readily available. The feature denoising approach proposed here is mostly similar to the SPLICE method of [19], which also requires clean and noisy speech for training (mentioned as stereo training data or parallel corpus), and like our methods does not assume noise stationarity. In fact, this method is very similar to the parallel training algorithm that we describe later. The main purpose of this paper, though, is to introduce our previously derived nonparallel training algorithm [4], [5] to the problem of speech enhancement [20]. The advantage of this method when compared to parallel training and SPLICE is the fact that there is no need for the clean and noisy speech to contain the same context. For this algorithm, initial conversion (estimation) parameters are obtained from a different speaker and noise characteristics pair, using a parallel corpus; these conversion parameters are then adapted to the speaker and noise characteristics of interest using nonparallel speech data (clean and noisy speech of the speaker of interest), through a parameter adaptation procedure similar to what is encountered in speech recognition. The training phase is simplified with this latter approach, since only few sentences of clean speech are needed, while the noisy speech is readily available. It is important to note that in this paper we employ a user-centric approach, i.e., the speech data we use for training come from the same speaker whose speech we attempt to enhance. In many scenarios, this is possible to implement in practice, while the results provided in this paper indicate that our methods can be easily generalized for the case when the data of multiple speakers are available but not necessarily of the particular speaker of interest.

It is also of interest to note that the method of [21] operates similarly to our IWF algorithm. In [21], the clean speech is estimated using minimum mean-squared error (mmse) estimation of the spectral envelope [by means of a trained Gaussian mixture model (GMM)], followed by Wiener filtering. This approach is in the same spirit as the IWF enhancement algorithm presented here, since in our work we also apply mmse estimation of the spectral envelope followed by Wiener filtering. The difference in our approach is that there is no model assumption for the noise (in [21] the noise is assumed to be Gaussian), which is achieved by assuming here a second GMM for the noisy speech.

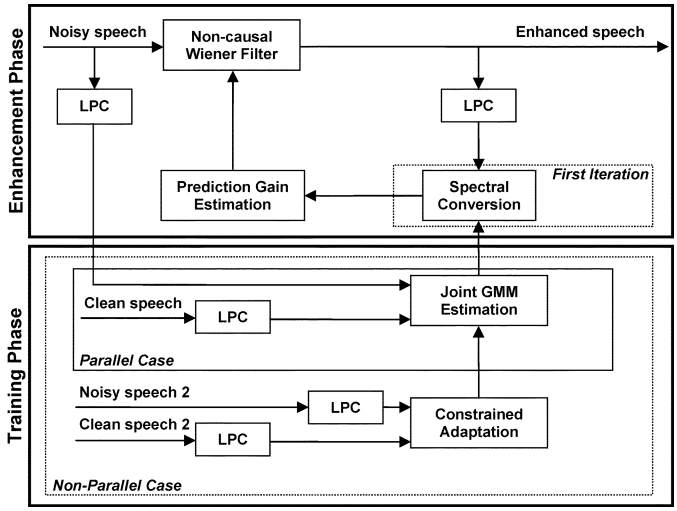

Fig. 1. Block diagram outlining spectral conversion for a parallel and nonparallel corpus within the IWF framework. Nonparallel training is achieved by adaptation of the parameters derived from parallel training of a different speaker and noise conditions.

In order to better demonstrate our approach, we concentrate in this paragraph our attention to the IWF algorithm, keeping in mind that the expectation-maximization iterative Kalman filter (KEMI) (also presented in the following sections) operates very similarly in philosophy with the advantage of being more suitable for colored and nonstationary noise. In Fig. 1, the block diagram of the proposed algorithms (original IWF and IWF using parallel and nonparallel spectral conversion) is given. The upper part of the diagram—excluding the spectral conversion block—corresponds to the original IWF. The noisy speech at each iteration is filtered with the noncausal Wiener filter, and from the enhanced signal the AR parameters [obtained using linear prediction (LPC)] are extracted to be used for the Wiener filter of the next iteration. At the first iteration, the noncausal Wiener filter is initialized with unity, meaning that the initial AR parameters of the clean speech are estimated directly from the noisy speech. The application of spectral conversion to the problem is shown in the diagram by the addition of the lower part denoted as "training phase." The upper box of the training phase part corresponds to the parallel conversion case, while the addition of the lower box corresponds to the nonparallel conversion. The assumption is that when spectral conversion is applied, the result is better estimation of the clean speech parameters rather than simply using the noisy speech parameters. 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 (words, sentences, etc.) must be available from the clean and noisy speech. This restriction is highly impractical in real-life scenarios for the problem of speech enhancement. In [4], [5], 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 [3]. For the particular pair of clean and noisy speech that we focus on, a nonparallel corpus is available for training. Constrained adaptation techniques allow for deriving the needed conversion parameters by relating the nonparallel corpus to the parallel corpus. We show that the speaker and noise characteristics in the two pairs of speech data can differ, not only in amplitude (SNR) but in spectral properties as well.

To summarize, in this paper we propose two mmse estimation methods for enhancing popular filtering algorithms for speech enhancement (the Wiener and Kalman filters). The mmse estimation methods are based on a speech corpus (used to train an estimation model), which in this paper is clean and noisy speech from the particular speaker whose speech must be enhanced. The noisy speech must correspond to the same conditions that are present during the enhancement phase. In one of the methods (parallel conversion), the speech and noisy speech data must contain the same speech context (parallel corpus), so that the spectral vectors of the noisy and clean speech can be time-aligned during training. The other mmse estimation method that is described is based on our previously derived nonparallel estimation method. In this method, clean and noisy speech from the particular speaker is still required, but they need not contain the same context (nonparallel corpus), which allows for a far more practical training procedure. The nonparallel estimation method operates by *adapting* the estimation parameters from a different speaker's noisy/clean speech parallel training data (referred to as the *initial* conversion pair), to the speaker whose speech we want to enhance. The nonparallel corpus is necessary exactly for performing this adaptation procedure. We note that for the initial conversion pair, not only the speaker but also the noise conditions can be different (the noise can be of different signal-to-noise ratio—but also of spectral content—than the noise that is actually present during the enhancement phase). However, the nonparallel corpus must still contain the noisy and clean speech from the particular speaker of interest and in the same noise conditions as those prevailing during the enhancement phase. As we show later, it is also possible to relax the requirement that the speech data come from the particular speaker (speaker-dependent enhancement), if a corpus that contains speech from several speakers is available.

The remainder of this paper is organized as follows. In Sections II and III, we briefly describe the IWF and KEMI algorithms for speech enhancement, respectively. In Section IV, we examine a popular algorithm for spectral conversion, which was found to be very suitable as a basis for our previously proposed nonparallel spectral conversion method [4], [5], described in Section V. In Section VI, simulation results are given for the IWF-based methods applied to white Gaussian noise (Section VI-A), and for the KEMI-based methods applied to colored nonstationary noise (Section VI-B). In Section VI-C, IWF- and KEMI- based methods are applied to speech in additive white noise, in order to provide a common ground for discussing their properties. Section VII concludes with a brief summary of the proposed approach.

II. 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 noncausal 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

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

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 is applied at each segment. We refer to such a segment-by-segment procedure as frame-wise processing, to distinguish it from a sample-by-sample procedure. For the speech enhancement algorithms that we use as a basis for our approach (i.e., IWF and KEMI), frame-wise processing is an important property since it is needed so that we can apply the spectral conversion methods as a preprocessing step (spectral conversion is inherently a frame-wise processing procedure as it can be seen in later sections).

For IWF, usually a small number of iterations for each segment is required for convergence, so the computational requirements of the algorithm are modest. 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.

III. Kalman Filter for Speech Enhancement

Again, we assume that $y(n)$ is the noisy signal, $s(n)$ is the clean speech signal, and $d(n)$ is the additive noise that is uncorrelated with $s(n)$. We follow the method of [17]. The algorithms that we describe operate successively in analysis segments (also denoted here as frames) of the signals (i.e., frame-wise processing, which is an important property as explained in the previous section). For each frame, the speech signal is assumed to follow an autoregressive (AR) model

$$s(n) = -\sum_{i=1}^{p} a_i s(n-i) + \sqrt{g_s} u(n) \tag{4}$$

where $u(n)$ is the excitation signal, assumed to be white noise with zero mean and unit variance, g_s is the spectral level, and a_i are the AR coefficients (order p). The noise is assumed as possibly nonwhite and more specifically to follow an AR model similar to (4)

$$d(n) = -\sum_{i=1}^{q} b_i d(n-i) + \sqrt{g_d} w(n) \tag{5}$$

with $w(n)$ as the zero-mean unit-variance white noise, g_d the noise spectral level, and b_i the AR coefficients (order q). These equations can be written in state-space form as

$$\begin{aligned} \boldsymbol{x}(n) &= \boldsymbol{\Phi}\boldsymbol{x}(n-1) + \mathbf{G}\boldsymbol{r}(n) \\ y(n) &= \boldsymbol{h}^T\boldsymbol{x}(n) \end{aligned} \tag{6}$$

where the state vector $\boldsymbol{x}(n)$ is given by

$$\begin{aligned} \boldsymbol{x}^T(n) &= \left[\boldsymbol{s}_{p-1}^T(n-1)s(n)\boldsymbol{d}_{q-1}^T(n-1)d(n)\right] \\ \boldsymbol{s}_p^T(n) &= [s(n-p+1)s(n-p+2)\cdots s(n)] \\ \boldsymbol{d}_q^T(n) &= [d(n-q+1)d(n-q+2)\cdots d(n)]. \end{aligned} \tag{7}$$

The state transition matrix $\boldsymbol{\Phi}$ can be easily found from the AR speech and noise models (4) and (5), and has a specific structure containing the AR coefficients of the speech and noise processes. Similarly, $\mathbf{G}$ is a matrix of specific structure containing $\sqrt{g_s}$ and $\sqrt{g_d}$, while $\boldsymbol{h}$ is the following vector:

$$\boldsymbol{h}^T = [\underbrace{0\,0\,\cdots}_{p-1}\,1\,\underbrace{0\,0\,\cdots}_{q-1}\,1] \tag{8}$$

and

$$\boldsymbol{r}(n) = [u(n)w(n)]^T. \tag{9}$$

If the parameters a_i, b_i, g_s, and g_d were known, then matrices $\boldsymbol{\Phi}$ and $\mathbf{G}$ would be known and the standard Kalman filter would be obtained, that provides the optimal mmse estimate of the state vector (and thus the clean speech signal). In practice, however, these parameters are not available. The KEMI algorithm of [17] estimates these parameters iteratively, within the Kalman filter algorithm. This approach is reviewed next.

The KEMI algorithm uses the expectation-maximization (EM) algorithm for iteratively estimating the speech and noise AR model parameters, applying the Kalman filter at each iteration. We use the notation

$$\begin{aligned} \boldsymbol{a} &= [a_p\ a_{p-1}\ \cdots\ a_1]^T \\ \boldsymbol{b} &= [b_q\ b_{q-1}\ \cdots\ b_1]^T \\ \boldsymbol{\theta} &= [\mathbf{a}^T\ g_s\ \boldsymbol{b}^T\ g_d]^T. \end{aligned} \tag{10}$$

Also, $\hat{\boldsymbol{\theta}}^{(l)}$ denotes the estimate of $\boldsymbol{\theta}$ after the lth iteration. We denote

$$\boldsymbol{y} = [y(1)\ y(2)\ \cdots\ y(N)]^T \tag{11}$$

as the vector of measurements for the current analysis frame. We denote as $\widehat{(\cdot)} = E_{\hat{\boldsymbol{\theta}}^{(l)}}(\cdot|\boldsymbol{y})$. To obtain the parameter estimate at iteration $l+1$, we use the following two-step EM procedure.

E-STEP: We denote the current state estimate and state covariance estimate respectively as

$$\begin{aligned} \boldsymbol{\mu}(n|N) &= \widehat{\boldsymbol{x}(n)} \\ \boldsymbol{P}(n|N) &= \widehat{\boldsymbol{x}(n)\boldsymbol{x}^T(n)} - \widehat{\boldsymbol{x}(n)}\,\widehat{\boldsymbol{x}(n)}^T. \end{aligned} \tag{12}$$

These can be found using the well-known Kalman filter recursion (propagation and updating equations), followed by the smoothing recursion. We omit the equations here, the interested reader is referred to [17]. The estimation equations are similar to the standard Kalman filter, with the difference that matrices $\boldsymbol{\Phi}$ and $\mathbf{G}$ are substituted by $\hat{\boldsymbol{\Phi}}$ and $\hat{\mathbf{G}}$ which are the matrices containing the current estimates of the AR parameters of the speech and noise processes (from the M-Step of the previous iteration), which is the reason that this iterative EM procedure is needed. The E-Step is followed by the M-Step providing the parameter estimates for the next iteration:

M-Step: The parameter estimates are given by

$$\begin{aligned} \hat{\boldsymbol{a}}^{(l+1)} &= -\left[\sum_{i=1}^{N}\boldsymbol{s}_p(n-1)\widehat{\boldsymbol{s}_p}(n-1)^T\right]^{-1}\sum_{i=1}^{N}\boldsymbol{s}_p(n\widehat{-1})s(n) \\ \hat{\boldsymbol{b}}^{(l+1)} &= -\left[\sum_{i=1}^{N}\boldsymbol{d}_q(n-1)\widehat{\boldsymbol{d}_q}(n-1)^T\right]^{-1}\sum_{i=1}^{N}\boldsymbol{d}_q(n\widehat{-1})d(n) \\ \hat{g}_s^{(l+1)} &= \frac{1}{N}\sum_{i=1}^{N}\left[\hat{s}^2(n) + \left(\hat{\boldsymbol{a}}^{(l+1)}\right)^T\boldsymbol{s}_p(n\widehat{-1})s(n)\right] \\ \hat{g}_d^{(l+1)} &= \frac{1}{N}\sum_{i=1}^{N}\left[\hat{d}^2(n) + \left(\hat{\boldsymbol{b}}^{(l+1)}\right)^T\boldsymbol{d}_q(n\widehat{-1})d(n)\right]. \end{aligned} \tag{13}$$

All the various estimates that are necessary in the aforementioned equations can be obtained as submatrices of $\widehat{\boldsymbol{x}(n)\boldsymbol{x}^T}(n) = \mathbf{P}(n|N) + \boldsymbol{\mu}(n|N)\boldsymbol{\mu}^T(n|N)$. It is of interest to note the similarity of the above equations with the Yule–Walker equations [22]. For the remainder of this paper, we use the delayed Kalman filter estimate (fixed-lag smoothing) for reducing the computational complexity of the algorithm. This means that we use $\hat{s}(n-p+1)$ as the current signal estimate (delay of $p-1$ samples), which is the first entry of $\boldsymbol{\mu}(n|N)$, and similarly for the noise estimate. The advantage of fixed-lag smoothing is that the smoothing equations need not be computed, which results in significantly fewer computations, while good performance is retained. Note that an initialization of the speech and noise AR parameters is required, which can be simply obtained from the noisy speech. Higher-order statistics can alternatively be used for the initialization [17]; in our experiments, this procedure did not offer any advantage and thus was not applied.

In the next two sections we provide an alternate approach to the initialization of the AR speech parameters needed in both IWF and KEMI algorithms. In Section IV, we present an estimation procedure of the clean speech AR parameters based on the noisy parameters, using a parallel training corpus, while in Section V, a similar procedure is applied, which does not require a parallel speech corpus.

IV. Spectral Conversion

In this section, we assume that training speech is available from a parallel corpus, which means that the training data contain same context clean and noisy speech waveforms. From these waveforms, we extract the parameters that model their short-term spectral properties [in this paper, we use the line spectral frequencies (LSFs) due to their desirable interpolation

properties [3]]. The LSFs are known to have a 1-1 correspondence with the AR spectral parameters that are needed in the IWF and KEMI algorithms. The result of the short-time analysis is a collection of two vector sequences, $[\boldsymbol{z_y}_1 \boldsymbol{z_y}_2 \ldots \boldsymbol{z_y}_n]$ and $[\boldsymbol{z_s}_1 \boldsymbol{z_s}_2 \ldots \boldsymbol{z_s}_n]$, of noisy and clean speech spectral vectors, respectively. The objective of spectral conversion methods is to derive a function $\mathcal{F}(\cdot)$ which, when applied to vector $\boldsymbol{z_y}_k$, produces a vector close in some sense to vector $\boldsymbol{z_s}_k$. A Gaussian mixture model (GMM) is often collectively represented as $\{p(\omega_i), \boldsymbol{\mu}_i^x, \boldsymbol{\Sigma}_i^{xx}\}$ where ω_i denotes a particular Gaussian class $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_i^x, \boldsymbol{\Sigma}_i^{xx})$ (i.e., a Gaussian pdf with mean $\boldsymbol{\mu}_i^x$ and covariance $\boldsymbol{\Sigma}_i^{xx}$). GMMs have been successfully applied to the voice conversion problem [2], [3]. GMMs approximate the unknown probability density function (pdf) of a random vector $\boldsymbol{z}$ as a mixture of Gaussians

$$\mathrm{g}(\boldsymbol{z}) = \sum_{i=1}^{M} p(\omega_i) \mathcal{N}\left(\boldsymbol{z}; \boldsymbol{\mu}_i^z, \boldsymbol{\Sigma}_i^{zz}\right) \tag{14}$$

where $p(\omega_i)$ is the prior probability of class ω_i, and $\mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The parameters of the GMM (mean vectors, covariance matrices, and prior probabilities of each Gaussian class), can be estimated from the observed data using the EM algorithm [23].

We focus on the spectral conversion method of [3], which offers great insight as to what the conversion parameters represent. Assuming that $\boldsymbol{z_y}$ and $\boldsymbol{z_s}$ are jointly Gaussian for each class ω_i, then, in mean-squared sense, the optimal choice for the function $\mathcal{F}$ is

$$\begin{aligned} \mathcal{F}\left(\boldsymbol{z_y}_k\right) &= \mathrm{E}\left(\boldsymbol{z_s} | \boldsymbol{z_y}_k\right) \\ &= \sum_{i=1}^{M} p\left(\omega_i | \boldsymbol{z_y}_k\right) \\ &\quad \times \left[\boldsymbol{\mu}_i^{z_s} + \boldsymbol{\Sigma}_i^{z_s z_y} {\boldsymbol{\Sigma}_i^{z_y z_y}}^{-1} \left(\boldsymbol{z_y}_k - \boldsymbol{\mu}_i^{z_y}\right)\right] \end{aligned} \tag{15}$$

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

$$p\left(\omega_i | \boldsymbol{z_y}_k\right) = \frac{p(\omega_i) \mathcal{N}\left(\boldsymbol{z_y}_k; \boldsymbol{\mu}_i^{z_y}, \boldsymbol{\Sigma}_i^{z_y z_y}\right)}{\sum_{j=1}^{M} p(\omega_j) \mathcal{N}\left(\boldsymbol{z_y}_k; \boldsymbol{\mu}_j^{z_y}, \boldsymbol{\Sigma}_j^{z_y z_y}\right)}. \tag{16}$$

All the parameters in the two above equations are estimated using the EM algorithm on the joint model of $\boldsymbol{z_y}$ and $\boldsymbol{z_s}$, i.e., $\boldsymbol{z_{ys}} = [\boldsymbol{z_y}^T \boldsymbol{z_s}^T]^T$ (where ^{T} denotes transposition). In practice, this means that the EM algorithm is performed during training on the concatenated vectors $\boldsymbol{z_y}_k$ and $\boldsymbol{z_s}_k$. A time-alignment procedure is required in this case, and this is only possible when a parallel corpus is used. For the speech enhancement problem, this translates into a need for the noisy speech training data to contain the same utterances (words, sentences, etc.) with the clean speech training data, which is prohibitive in practice. The covariance matrices $\boldsymbol{\Sigma}_i^{z_y z_y}$, $\boldsymbol{\Sigma}_i^{z_s z_y}$ and the means $\boldsymbol{\mu}_i^{z_y}$, $\boldsymbol{\mu}_i^{z_s}$ in (15) and (16) can be directly obtained from the estimated covariance matrices and means of $\boldsymbol{z_{ys}}$, since

$$\boldsymbol{\Sigma}_i^{z_{ys} z_{ys}} = \begin{bmatrix} \boldsymbol{\Sigma}_i^{z_y z_y} & \boldsymbol{\Sigma}_i^{z_y z_s} \\ \boldsymbol{\Sigma}_i^{z_s z_y} & \boldsymbol{\Sigma}_i^{z_s z_s} \end{bmatrix} 890 \quad \boldsymbol{\mu}_i^{z_{ys}} = \begin{bmatrix} \boldsymbol{\mu}_i^{z_y} \\ \boldsymbol{\mu}_i^{z_s} \end{bmatrix}. \tag{17}$$

Another issue is that performance considerations, when using the adaptation procedure described in the next section, 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 [24].

V. Constrained GMM Estimation

In the previous section, we described a 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 and KEMI algorithms during the first iteration. However, a parallel training corpus will be required in this case, which as explained is impractical to acquire for the speech enhancement problem. 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 nonparallel corpus. We show that this is possible under the assumption that a parallel speech corpus is available for a *different* noisy and clean speech pair (i.e., different speaker and noise conditions). In order to achieve this result, we apply the maximum-likelihood constrained adaptation method [25], which offers the advantage of a simple probabilistic linear transformation leading to a mathematically tractable solution.

We 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 nonparallel corpus exists. From the parallel corpus, we obtain a joint GMM model, derived as explained in Section IV. The spectral vectors that correspond to the noisy speech are considered as realizations of random vector $\boldsymbol{z_y}$, while $\boldsymbol{z_s}$ corresponds to the clean speech of the parallel corpus. From the nonparallel corpus, we also obtain a sequence of spectral vectors, considered as realizations of random vector $\boldsymbol{z_y}'$ for the noisy speech and $\boldsymbol{z_s}'$ for the clean speech. We then relate the random variables $\boldsymbol{z_y}'$ and $\boldsymbol{z_y}$, as well as $\boldsymbol{z_s}'$ and $\boldsymbol{z_s}$, in order to derive a conversion function for the nonparallel corpus based on the parallel corpus parameters.

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

$$\boldsymbol{z_y}' = \mathbf{A}_j \boldsymbol{z_y} + \boldsymbol{b}_j \text{ with probability } p(\lambda_j | \omega_i), \quad j = 1, \ldots, N. \tag{18}$$

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

$$\sum_{j=1}^{N} p(\lambda_j | \omega_i) = 1, \quad i = 1, \ldots, M. \tag{19}$$

In the aforementioned 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 $\boldsymbol{z_y}$), and $\boldsymbol{b}_j$ is a vector of the same dimension with $\boldsymbol{z_y}$. The clean speech random (spectral) vectors $\boldsymbol{z_s}'$ and $\boldsymbol{z_s}$ are related by another probabilistic linear transformation, similar to (18), 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 are the same for $\boldsymbol{z_y}$ and $\boldsymbol{z_s}$ by design in Section IV. All the unknown parameters can be estimated by use of the nonparallel 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 [4], [5], that the conversion function for the nonparallel case is

$$\begin{aligned}\mathcal{F}(\boldsymbol{z}_{\boldsymbol{y}_k'}) &= \mathrm{E}(\boldsymbol{z}_{\boldsymbol{s}}'|\boldsymbol{z}_{\boldsymbol{y}_k'}) \\ &= \sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{\rho=1}^{L} p(\omega_i|\boldsymbol{z}_{\boldsymbol{y}_k'})p(\lambda_j|\boldsymbol{z}_{\boldsymbol{y}_k'},\omega_i) \\ &\quad \times p(\kappa_\rho|\omega_i)\Big[\boldsymbol{C}_\rho\boldsymbol{\mu}_i^{z_s} + \boldsymbol{d}_\rho + \boldsymbol{C}_\rho\boldsymbol{\Sigma}_i^{z_s z_y}{\boldsymbol{\Sigma}_i^{z_y z_y}}^{-1} \\ &\quad \times \boldsymbol{A}_j^{-1}\big(\boldsymbol{z}_{\boldsymbol{y}_k'} - \boldsymbol{A}_j\boldsymbol{\mu}_i^{z_y} - \boldsymbol{b}_j\big)\Big] \end{aligned} \tag{20}$$

$$p\left(\omega_i|\boldsymbol{z}_{\boldsymbol{y}_k'}\right) = \frac{p(\omega_i)\sum_{j=1}^{N} p(\lambda_j|\omega_i)\mathrm{g}\left(\boldsymbol{z}_{\boldsymbol{y}_k'}|\omega_i,\lambda_j\right)}{\sum_{i=1}^{M}\sum_{j=1}^{N} p(\omega_i)p(\lambda_j|\omega_i)\mathrm{g}\left(\boldsymbol{z}_{\boldsymbol{y}_k'}|\omega_i,\lambda_j\right)} \tag{21}$$

$$p\left(\lambda_j|\boldsymbol{z}_{\boldsymbol{y}_k'},\omega_i\right) = \frac{p(\lambda_j|\omega_i)\mathrm{g}\left(\boldsymbol{z}_{\boldsymbol{y}_k'}|\omega_i,\lambda_j\right)}{\sum_{j=1}^{N} p(\lambda_j|\omega_i)\mathrm{g}\left(\boldsymbol{z}_{\boldsymbol{y}_k'}|\omega_i,\lambda_j\right)}, \tag{22}$$

$$\mathrm{g}\left(\boldsymbol{z}_{\boldsymbol{y}}'|\omega_i,\lambda_j\right) = \mathcal{N}\left(\boldsymbol{z}_{\boldsymbol{y}}';\mathbf{A}_j\boldsymbol{\mu}_i^{z_y} + \boldsymbol{b}_j, \mathbf{A}_j\boldsymbol{\Sigma}_i^{z_y z_y}\mathbf{A}_j^T\right). \tag{23}$$

VI. Simulation Results

In this section, we test the performance of the parallel and nonparallel spectral conversion methods described in the previous paragraphs to the speech enhancement problem within the IWF (Section VI-A) and KEMI (Section VI-B) frameworks. The IWF-based algorithm is tested using white noise, since this algorithm is designed for this type of noise, while KEMI is tested using colored noise (car interior noise) with a low degree of nonstationarity. In Section VI-C, we apply both IWF- and KEMI- based methods to speech in additive white noise, in order to discuss their properties regarding the quality of the enhanced signals.

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 [26].[1] This is a parallel corpus and is used for both the parallel and nonparallel training cases that are examined in this section, in a manner explained in the next paragraphs.

A. IWF Results

In this section, we test the IWF-based methods using additive white noise. We use 40-ms windows (the sampling rate is 22.050 kHz) and the spectral vectors used here are the LSF's (28th order) due to their favorable interpolation properties. For these experiments, we use white Gaussian noise. We test the performance of the two conversion algorithms proposed here (one case (15) for parallel training and one (20) for nonparallel training), in comparison to the unconstrained IWF and spectral subtraction [6]. The ideal error for the IWF method is given as well, i.e., using the all-pole coefficients of the clean speech signal, which are available only in the simulation environment. This is the ideal case for the original IWF as well as the two conversion-based methods; thus, it is expected to give the maximum performance that can be achieved with all three approaches. It is important to note that the corpus used contains a total of 50 sentences, of which a total of 40 are used for training purposes (as explained next) and the remaining ten are used for testing. All the results given in this section are averaged over these ten sentences and, in addition, for each sentence the result is the average of ten different realizations of noise.

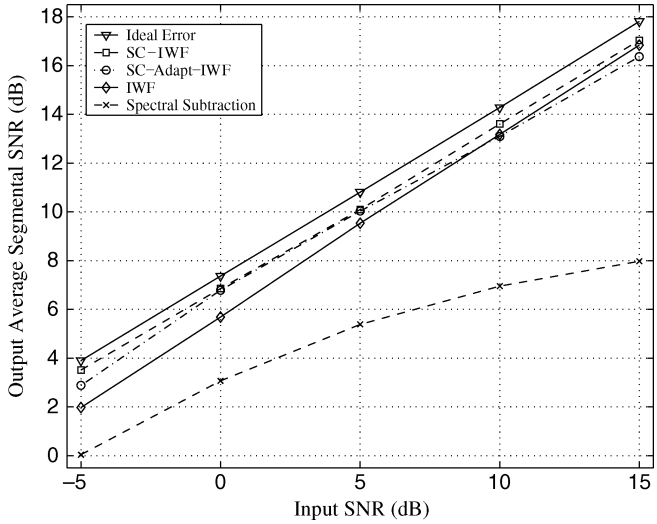

Fig. 2. Resulting ASSNR (dB) for different values of input SNR (white noise), for the five cases tested, i.e., perfect prediction (ideal error), the iterative Wiener filter (IWF), spectral conversion for IWF (SC-IWF, parallel corpus), spectral conversion by adaptation for IWF (SC-Adapt-IWF, nonparallel 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 nonparallel training (SC-Adapt-IWF)], compared with the IWF algorithm, spectral subtraction, and the theoretically best possible performance of the conversion-enhanced IWF (i.e., using the original AR parameters from the clean speech signal). For SC-IWF, the number of GMM parameters for training is 16 and the number of vectors in training is 5000, 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 5000. From the figure it is evident that the SC-IWF algorithm improves on the IWF algorithm, especially in low input SNRs, which is exactly what is desired. In many cases in our simulations the performance improvement reached 2 dB, which is quite important perceptually in low SNRs. 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 per-

[1][Online] Available: http://www.cslu.ogi.edu/corpora/voices/

TABLE I
RESULTING ASSNR (DECIBELS) FOR INPUT SNR OF 0 dB (WHITE NOISE) FOR ITERATIVE WIENER FILTER (IWF), PERFECT PREDICTION (IDEAL ERROR), SPECTRAL SUBTRACTION, SPECTRAL CONVERSION WITH IWF (SC-IWF), AND SPECTRAL CONVERSION FOLLOWED BY ADAPTATION AND IWF (SC-ADAPT-IWF)

Method	ASSNR
IWF	5.6839
Ideal Error	7.3654
Spectral Subtraction	3.0678
SC-IWF	6.8538
SC-Adapt-IWF	6.7877

TABLE II
RESULTING ASSNR IN DECIBELS (IWF WITH PARALLEL TRAINING, 0 dB INPUT SNR, WHITE NOISE), FOR DIFFERENT NUMBERS OF GMM PARAMETERS (FOR 5000 VECTORS) AND TRAINING VECTORS (FOR 16 GMM PARAMETERS)

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 III
RESULTING ASSNR IN DECIBELS (IWF WITH NONPARALLEL TRAINING, 0 dB INPUT SNR, WHITE NOISE), FOR DIFFERENT NUMBERS OF ADAPTATION PARAMETERS (FOR 5,000 VECTORS) AND TRAINING VECTORS (FOR FOUR ADAPTATION PARAMETERS)

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

formance 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 SNRs, when it performs only marginally worse than SC-IWF. In higher SNRs the training corpus, parallel or nonparallel, 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 I for comparison with the results in Tables II and III.

In Table II, 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 5000. 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 under favorable listening conditions. 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 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 2000 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 III, the ASSNR is given for the nonparallel case and input SNR of 0 dB, for various choices of adaptation parameters (again, in (20) $L = N$) and training dataset size. When varying the number of adaptation parameters, the training dataset contains 5000 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., different context from different speaker and noise conditions, for which a parallel corpus is used with 16 GMM parameters and 5000 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 nonparallel case. It is evident that adaptation is indeed useful, 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. We also notice in the table that the result for adaptation of 0 parameters (no adaptation), while worse than what we obtain when using adaptation, it is nevertheless improved when compared to the results of the original IWF algorithm. This is an indication that the conversion-based algorithms proposed here can be easily generalized to the case when clean speech data of the particular speaker might not be available. In that case, speech from a different speaker from the corpus could be used and still result in improvement over IWF. This issue is more evident in the following section where KEMI results are discussed.

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 noteworthy 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 preprocessing step.

B. Kalman Filter Results

In this section, we measure the performance of our two proposed conversion algorithms (parallel and nonparallel conversion) as an improvement to the Kalman filter for speech enhancement. We use again the *VOICES* corpus. The background noise, added artificially to the speech signals, is car interior noise (with constant acceleration) obtained from the NOISEX-92 corpus [27]. This type of noise is colored with a

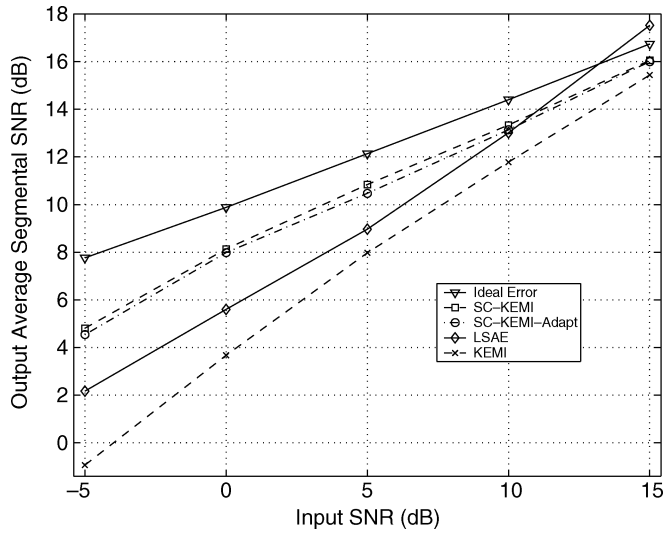

Fig. 3. ASSNR (decibels) for different values of input SNR (car noise), for the five cases tested, i.e., perfect prediction (ideal error), KEMI, spectral conversion followed by KEMI (SC-KEMI, parallel corpus), spectral conversion by adaptation followed by KEMI (SC-KEMI-Adapt, nonparallel corpus), and LSAE.

low degree of nonstationarity. The noise and speech signals were downsampled to 8 kHz for reducing the implementation demands of the various methods. We implemented and tested, in addition to our two proposed algorithms, the original KEMI algorithm of [17], as well as the LSAE algorithm of [12], for comparison. The latter has been shown to exhibit very desirable performance in [17] compared to the KEMI algorithm in output SNR sense.

In our implementation, we use a 32-ms analysis frame and (for the Kalman-based methods) LSF vectors of 22nd order for the speech signal (12th for the noise). The noise parameters were initialized (noise estimation for LSAE) using very few signal segments that did not contain any speech (initial segments of each recording). The error measure employed is again the output average segmental SNR. We test the performance of the algorithms using the ASSNR for various values of input (global) SNR. We test the performance of the two algorithms proposed here [one case (15) for parallel training and one (20) for nonparallel training], in comparison to the original KEMI algorithm and LSAE. The ideal error for both our methods (the desired LSFs with zero prediction error, only available in the simulation environment) is also given. As previously, from the 50 sentences of the corpus we use a total of 40 for training purposes (as explained next) and the remaining ten for testing. All the results given in this section are averaged over these ten sentences (with different noise segments added to each sentence).

In Fig. 3, the ASSNR is given for the five cases tested, for various values of input SNR. The five cases are: the two proposed algorithms for parallel and nonparallel training as an initialization to the KEMI algorithm (SC-KEMI and SC-KEMI-Adapt, respectively), the KEMI algorithm (iterative Kalman filter), the log-spectral amplitude estimation (LSAE) algorithm, as well as the theoretically best possible performance of the conversion-based approaches (the desired LSFs with zero prediction error are used for the initialization of KEMI). It is important to mention that the results for both our methods, as well as their ideal error performance, were obtained without use of the iterative Kalman procedure. In other words, the results were obtained by LSF estimation followed by the standard Kalman filter. We found that further iterations did not offer any significant improvement. For the KEMI algorithm, we obtained good results after 15 iterations. For the results in Fig. 3, we used around 20 000 training LSF vectors, which correspond to 40 sentences of the corpus. Later in this section, we discuss the effect of the size of the training corpus to the final results. Also, the number of (diagonal) GMM classes used for both the parallel and nonparallel methods is 16 [$M = 16$ in (15) and (20)], while the number of adaptation parameters is 4 for both the source and target speech [$L = N = 4$ in (20)]. For this figure, we plot the performance of the SC-KEMI-Adapt algorithm based on adaptation of the GMM conversion parameters of a different speaker from our corpus, in car interior noise of 10-dB SNR (i.e., the SNR is accurate only for the 10-dB input SNR case). From Fig. 3, we can see that the improvement in the KEMI algorithm using both the methods proposed in this paper is significant, especially for low input SNRs. For input SNR of -5 dB for example, the improvement is almost 6 dB for both methods, which is important perceptually. A very interesting observation is that the adaptation algorithm performs almost as well as the parallel algorithm. This was not expected, given that we have previously explained (for voice conversion) that adaptation will always perform worse than the parallel method since in parallel training we exploit an additional property of the corpus in an explicit manner. In [4], [5], we have shown that the variations in the estimation error are small between these two algorithms when compared to the distance between the initial and desired parameters. We can conclude that the Kalman filter does not exhibit much sensitivity to the small variations in the estimation error for the initialization parameters in contrast to the case of large estimation errors that are encountered in the original KEMI algorithm (i.e., estimating the clean parameters directly from the noisy speech). This is also encountered later in this section, when comparing the ASSNR when fine-tuning the GMM and adaptation parameters (Tables V and VI). In high input SNRs, the algorithms perform similarly (with the LSAE resulting in the best estimation results for 15-dB SNR), which is sensible since in high SNRs the speech initial parameters estimation from the noisy speech is very close to the desired. The results for input SNR of 0 dB are also given in Table IV for convenience.

TABLE IV
ASSNR (DECIBELS) FOR INPUT SNR OF 0 dB (CAR NOISE) FOR KEMI, PERFECT PREDICTION (IDEAL ERROR), LSAE, SPECTRAL CONVERSION AS AN INITIALIZATION TO KEMI (SC-KEMI), AND SPECTRAL CONVERSION FOLLOWED BY ADAPTATION FOLLOWED BY KEMI (SC-KEMI-ADAPT)

Method	ASSNR
KEMI	3.6702
Ideal Error	9.8798
LSAE	5.6004
SC-KEMI	8.1329
SC-KEMI-Adapt	7.9836

In Table V, the ASSNR is given for the parallel case (SC-KEMI) for 0-dB input SNR, for various numbers of GMM

TABLE V
RESULTING ASSNR IN DECIBELS (KEMI WITH PARALLEL TRAINING, 0 dB INPUT SNR, CAR NOISE), FOR DIFFERENT NUMBERS OF GMM PARAMETERS (FOR 20 000 VECTORS) AND TRAINING VECTORS (FOR 16 GMM PARAMETERS)

GMM's	2	4	8	16	32
ASSNR	8.1769	8.1338	8.1564	8.1329	8.0073
Vectors	500	1000	2000	5000	20000
ASSNR	7.9333	7.9784	8.0715	8.0611	8.1329

TABLE VI
RESULTING ASSNR IN DECIBELS (KEMI WITH NONPARALLEL TRAINING, 0 dB INPUT SNR, CAR NOISE), FOR DIFFERENT NUMBERS OF ADAPTATION PARAMETERS (FOR 20 000 VECTORS) AND TRAINING VECTORS (FOR FOUR ADAPTATION PARAMETERS)

Param.	0	1	2	4	6
ASSNR	0.3359	8.1757	8.0450	8.2340	8.0674
Vectors	500	1000	2000	5000	20000
ASSNR	7.7370	8.1359	7.9530	8.2912	8.2340

parameters and training vectors. When comparing the performance of the various numbers of GMM parameters, the vectors in training are 20 000. The number of GMM parameters does not seem to have an influence on the performance of the algorithm. For the second case examined in this table, namely the effect of the training dataset size on the algorithm performance, we use 16 GMM parameters. We can see that the performance of the algorithm improves slightly when more training vectors are available. The fact that only a small number of training data results in major improvement over KEMI is important, since this corresponds to requiring only a small amount of clean speech data. The fact that we have such a noteworthy improvement in the KEMI algorithm without large variations in the number of GMM parameters or training data is consistent with our previous observation (when comparing parallel versus nonparallel training), that the Kalman filter is not influenced much by small variations in the LSF estimation error.

In Table VI, the ASSNR is given for the nonparallel case (SC-KEMI-Adapt) and input SNR of 0 dB, for various choices of adaptation parameters [$L = N$ in (20)] and training dataset. When varying the number of adaptation parameters, the training dataset contains 20 000 vectors, and when varying the number of vectors in the training dataset, the number of adaptation parameters is $L = N = 4$. For the results in this table, the noise conditions of the parallel (*initial*) pair (i.e., initial conversion parameters) were obtained for *white noise* of 10-dB SNR. This choice was made so that we can show more evidently the effect of adaptation on the algorithm performance, since in this case the initial error (i.e., with no adaptation) is much larger than in the case when the initial pair contains the same type (car interior) noise. With no adaptation, i.e., simply applying the GMM parameters of a different speaker/noise pair to the speaker in car noise environment, the ASSNR is only 0.3359, which is worse than the original KEMI results for 0-dB SNR (3.6702 to be specific). On the other hand, we observe once again the lack of sensitivity of the Kalman filter to small LSF estimation errors (as long as the adaptation procedure is employed). We also observe that, similarly to the parallel case of Table V, increasing the number of training vectors consistently improves the algorithm performance, although not significantly. The fact that a small number of training data results in good algorithm performance is very positive, since in many cases gathering large numbers of data is impractical. We also mention at this point that the result for no adaptation when the initial conversion parameters are estimated from a different speaker/noise pair was measured to be 7.8166, when the training noise was car noise of 10-dB SNR, (i.e., training noise similar but of different SNR than the actual noise of 0 dB). This is of interest since this result is much improved when compared to the original KEMI. This observation justifies the claim that the conversion-based algorithms can be generalized to the case when clean speech from the particular speaker to be enhanced is not available, so that speech from a different speaker is used. This claim seems to hold when the noise in the training corpus is similar (but not necessarily of the same SNR) as the noise in the testing data.

1) Noise Estimation: In both KEMI and LSAE algorithms the noise power spectral density (PSD) is needed *a priori*, and is used in order to produce the current segment's clean speech estimate. Thus, there is a need to estimate the noise PSD on a segment-by-segment basis (every few milliseconds). In the results given so far in this section, the noise PSD was obtained from the first segment of the noisy speech signal, which is known that it contains only noise (speech silence). In other words, the noise estimate is accurate but at the same time it is not updated again for the duration of each sentence (in the order of a few seconds). This was chosen so that the results obtained can be considered accurate when compared to the practical scenario that the noise is estimated from the noisy speech. In this subsection, we are interested to show that indeed this is the case, and the results would be similar if we had used a practical method for noise estimation.

For achieving noise estimation in practice, two approaches are mostly popular. One is to use a voice activity detector (VAD), so that noise can be obtained from segments that are identified as silent. The problem with such approaches is that a false decision of the VAD will result in an inaccurate estimate of the noise. The alternative is to use soft-decision methods, when the noise estimation is not so much affected by a decision of whether the current segment of the noisy waveform contains noisy speech or noise only. One such method is the minimum statistics method of [28], where the noise estimation is based on tracking the minimum of the noise PSD. This method has been shown to result in very good performance compared to VAD estimation methods, and as such, has been incorporated for the results given in Table VII, for -5-dB SNR car noise. This value of input SNR is the lowest in our experiments and was chosen since in lower SNRs the effect of noise estimation in speech enhancement algorithms is more evident. This method is straightforward to use in conjunction with LSAE, but it can also be used in conjunction with any other method of speech enhancement that requires a noise estimate as part of the algorithm functionality. In this sense, we have also applied the minimum statistics method within the KEMI framework, for estimating the noise spectral envelope and the noise variance that is needed. The results of Table VII show the achieved ASSNR for LSAE and the KEMI-based methods (i.e., for the colored noise case). The results that correspond to the incorporation of

TABLE VII
ASSNR (DECIBELS) FOR INPUT SNR OF −5 dB (CAR NOISE), USING THE NOISE ESTIMATION METHOD OF [28] (COLUMN "WITH"), FOR ITERATIVE KALMAN FILTER (KEMI), PERFECT PREDICTION (IDEAL ERROR), LOG-SPECTRAL AMPLITUDE ESTIMATOR (LSAE), SPECTRAL CONVERSION AS AN INITIALIZATION TO KEMI (SC-KEMI), AND SPECTRAL CONVERSION FOLLOWED BY ADAPTATION FOLLOWED BY KEMI (SC-KEMI-ADAPT).

Enhancement Method	ASSNR	
	With	W/out
KEMI	0.2807	-0.9360
Ideal Error	6.7018	7.7643
LSAE	2.6394	2.1704
SC-KEMI	5.2802	4.8088
SC-KEMI-Adapt	5.1127	4.5416

noise estimation to the previously mentioned speech enhancement methods are given in the column denoted as "With" (i.e., with noise estimation). The column denoted as "Without" corresponds to the use of the first segment of the noisy signal, and are the same as the ones of Fig. 3 (given here for comparison). From the results of the table we can conclude that indeed the noise estimation does not change in a noticeable degree the results obtained in the previous paragraphs. The exception is the original KEMI algorithm (ASSNR from −0.9360 to 0.2807), which is still much lower than the rest of the methods described, and is also a trend that was not confirmed for other values of input SNR. For the remaining methods, we can see that the relative performance is very similar, and thus the conclusions in the previous paragraphs regarding the relative SNR results for the (parallel and nonparallel) conversion-based approaches compared to KEMI and LSAE are valid.

2) Listening Test: We conducted a listening test in order to judge the subjective quality of the enhanced signals using various of the methods described here for speech enhancement. For this test, we were interested to test the enhancement of speech under the car noise environment in −5-dB SNR. Thus, we tested all the methods that were implemented in this paper for the car noise environment, i.e., the KEMI-based methods as well as LSAE. Additionally, we used the noise estimation method that was applied in the previous paragraph for the results of Table VII. In the listening test, 15 volunteers participated, and we used three audio signals from our testing dataset, to which car noise was added. Each of the five enhancement methods was applied to the three noisy signals (referred to as Signals 1–3 in this section), resulting in a total of 15 enhanced signals. The listening test employed was a degradation category rating (DCR) test [29], in which each subject is presented (using high-quality headphones) each of the enhanced signals and the corresponding clean speech signal, and is asked to grade them using grades 1 to 5. These grades correspond to: 5 to "No quality degradation perceived" (compared to the clean speech signal), 4 to "Quality degradation perceived but not annoying," 3 to "Quality degradation perceived and is slightly annoying," 2 to "Quality degradation perceived and is annoying," and 1 to "Quality degradation perceived and is very annoying."

The DCR results are given in Fig. 4. From these results we can see that regarding the KEMI-based methods, the subjective results are consistent with the objective results of Table VII. In other words, for the KEMI-based methods, the ideal conversion for KEMI results in best enhancement, followed by parallel conversion, and in turn followed by the nonparallel conversion,

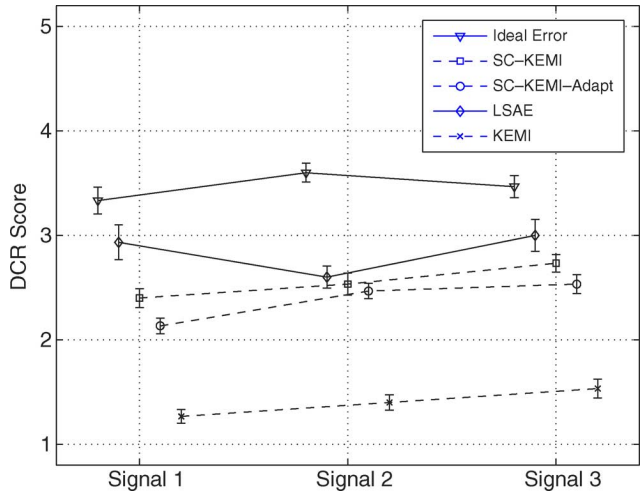

Fig. 4. Results from the DCR listening test, for input SNR of −5 dB (car noise), using the noise estimation method of [28], for KEMI, perfect prediction (ideal error), LSAE, spectral conversion as an initialization to KEMI (SC-KEMI), and spectral conversion followed by adaptation followed by KEMI (SC-KEMI-Adapt).

while the original KEMI was always ranked as the lowest in quality. It is interesting to note that for the subjective results, as in the objective results, parallel and nonparallel conversion methods perform very similarly, which is important given the practical advantages of nonparallel conversion. We also note the high-quality performance of LSAE as shown in Fig. 4. This might seem contradictory when compared to the objective results of Table VII, since objectively LSAE was shown to perform worse than the parallel and nonparallel conversion methods. However, this can be attributed to the fact that the lower SNR results of LSAE are due to low-frequency residual noise which is not so audible, while the residual noise of parallel and nonparallel methods was found to be more equally distributed along the frequency domain. Equally important is the fact that the KEMI-based methods seem to result in degradation of the high-frequency components of the enhanced signal, in contrast to LSAE. This issue is further discussed in the following section, and is analyzed using spectrograms of the enhanced signals. Due to these issues, LSAE was ranked second best (following the ideal conversion case) in the subjective tests, although the output SNR for this methods was in fact lower than the conversion-based enhancement methods. It is noteworthy that the ideal conversion method performed a great degree higher than the other enhancement methods both objectively and subjectively; this is an indication that methods aiming at improving the AR parameters of the clean speech from the noisy speech, such as the proposed conversion methods, are indeed very promising for the speech enhancement problem.

C. Discussion

In this section, our objective is to give an estimate of the quality of the speech signals that result from the enhancement algorithms proposed in this paper, in addition to the listening test of the previous section. In this section, we give examples of the resulting speech signals using spectrograms that allow us to more deeply evaluate the performance of the various algorithms as opposed to only examining the resulting SNR.

In order to judge the various methods in the same conditions, in this section we give results for speech corrupted by white Gaussian noise in 0-dB SNR. The speech signal used is one sentence from our corpus "The angry boy answered but didn't look up," used only for testing, downsampled at 8 kHz. The noisy speech signal was recorded as well (with artificially added noise), so that the exact same noisy signal was used for all algorithms. For both IWF- and KEMI-based algorithms the LSF order for the speech signals was 22, the window analysis was obtained using 64-ms segments, for parallel conversion 16-class GMMs were trained, and for nonparallel conversion 4 adaptation parameters ($L = N = 4$) were used. Regarding the corpus used, again the VOICES corpus was employed, using 15 of the total 40 training sentences for training the parallel conversion pairs, and another 15 sentences for the nonparallel adaptation (different sentences than those in parallel training). The methods examined in this section are the original IWF and KEMI algorithms, and their conversion-based improvements (with parallel and nonparallel conversion), including the ideal case of "perfect prediction" (i.e., using the clean speech AR parameters). In Table VIII, the various methods are ranked based on the resulting segmental SNR. From the table, we can see that KEMI performs better than IWF for white noise when enhanced by the conversion step, but in general the results are very close. This trend was maintained when we obtained results by averaging more testing data. However, it is interesting to note that the perfect prediction case for KEMI produced significantly better results than the perfect prediction for IWF, which is a motivation for considering KEMI as a more viable alternative for future research. In this sense, it is of interest to note that the IWF results in this section were obtained—as in previous sections—using the ideal number of iterations, which is not possible in practice. Thus, compared to IWF, KEMI exhibits a more robust behavior, while on the other hand KEMI is more computationally demanding. Finally, note that for the white noise results in this section, KEMI required about ten iterations for best performance, including the conversion-based approaches, while the AR order for the noise was set to 0, i.e., the noise was assumed to be white in the original model. The increased number of iterations in the conversion-based methods was found to be needed for better estimating the clean speech signal power, which is under investigation as to why this was important for white and not for colored noise.

In Fig. 5, spectrograms are given for the methods mentioned in the previous paragraph, corresponding to the SNR results of Table VIII. For more clearly showing the spectrogram details (given the space constraints), only the first part "The angry boy answered" of the sentence is shown in the figure. From the spectrograms it can be seen that for the ideal case for both KEMI and IWF, the resulting speech quality is very good while the noise is clearly diminished. It is apparent from the figure that the ideal KEMI case performs better then the corresponding IWF, as the results in Table VIII indicate. From the table, we also see that the parallel conversion KEMI method produces better resulting ASSNR than the ideal IWF case; however, from the figure, we can see that this happens at the expense of the resulting quality, since higher frequency components are degraded for the former method. In this sense, we can also see that the parallel conversion results for both KEMI and IWF produce better quality speech when compared to the corresponding nonparallel variants, for which the frequency components above 1000 Hz are severely diminished. This is an issue that was not apparent when comparing the resulting ASSNRs of the various methods. Finally, we note that for all methods (including the ideal conversion cases), unvoiced speech is degraded, and this can be easily seen from the spectrograms. We note that the observations of this section are in line with—and help us gain better insight regarding—what the listeners observed during the DCR listening test.

TABLE VIII
RESULTING ASSNR FOR THE VARIOUS IWF- AND KEMI- BASED ALGORITHMS PROPOSED IN THE PAPER. "IDEAL" CORRESPONDS TO THE IDEAL CONVERSION CASE (WHEN USING THE CLEAN SPEECH PARAMETERS), "PARAL." CORRESPONDS TO PARALLEL CONVERSION, AND "ADAPT." CORRESPONDS TO NONPARALLEL CONVERSION. THE ADDITIVE NOISE IS WHITE IN 0 dB SNR. THE VARIOUS METHODS ARE RANKED BASED ON THE RESULTING ASSNR AND CORRESPOND TO THE SPEECH SIGNALS IN FIG. 5

Method	ASSNR
KEMI-Ideal	5.7975
KEMI-Paral.	4.4249
IWF-Ideal	4.3253
KEMI-Adapt.	4.0200
IWF-Paral.	3.7934
IWF-Adapt.	3.3519
IWF	2.8055
KEMI	2.7046

As a concluding remark for this section, we mention that KEMI-based methods show more promise when compared to the IWF-based methods, which was mainly shown when comparing the ideal prediction cases. On the other hand, complexity for KEMI-based methods remains an important issue. Regarding quality, it is apparent that the better the AR parameters estimation, the better speech quality we will obtain in the enhanced signal. Even when the output SNR drops, if the AR parameter estimation is low in accuracy (which is more evident in nonparallel conversion), then the quality of the enhanced signal will be degraded, especially regarding high-frequency components.

VII. CONCLUSION

For single-channel speech enhancement, numerous algorithms have been proposed. Two of the most successful approaches are based on linear filtering techniques, more specifically the Wiener and Kalman filters. On the other hand, for many practical scenarios it is possible to have prior access to clean speech signals, and for that case a different class of enhancement algorithms have been proposed. In this paper, we attempt to combine the advantages of linear filters regarding their performance and the good signal quality they produce, with the additional prior information that is often available in practice. Our approach has been to provide initial estimates of the clean speech parameters from the noisy speech, using spectral conversion. In order to provide a practically useful algorithm, we introduced our previously derived nonparallel conversion method, which estimates the clean speech features from the noisy features with the use of a small training clean speech corpus. In the nonparallel conversion method, the clean and noisy speech data that are required need not contain the

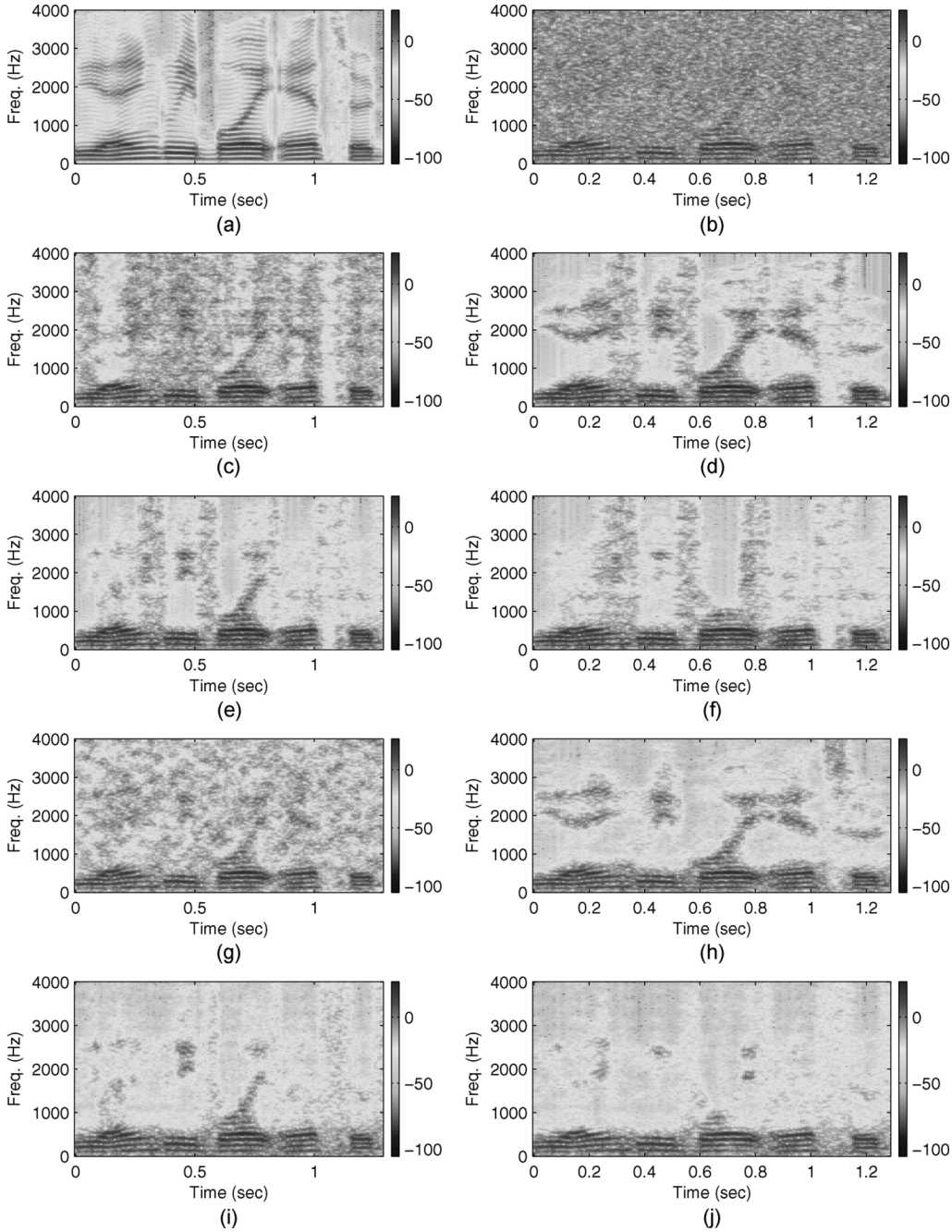

Fig. 5. Spectrograms of (a) the clean speech signal "The angry boy answered," (b) the noisy speech with 0 dB SNR, and the enhanced speech processed by (c) the IWF algorithm, (d) IWF preceded by perfect prediction (ideal case), (e) IWF preceded by parallel conversion, (f) IWF preceded by nonparallel conversion, (g) the KEMI algorithm, (h) KEMI preceded by perfect prediction (ideal case), (i) KEMI preceded by parallel conversion, (j) KEMI preceded by nonparallel conversion.

same context, and thus the data collection process is greatly simplified. The results provided in this paper indicate that the proposed nonparallel conversion method performs almost as well as parallel conversion, both objectively and subjectively, which is important given the practical advantages of nonparallel conversion. At the same time, we showed that application of voice conversion as a first step to speech enhancement algorithms that are based on the clean speech AR parameters produces a major improvement as opposed to simply using the noisy AR parameters. In this sense, the conversion step presented here as part of IWF and KEMI algorithms can be applied in a wider context, whenever such a speaker-dependent approach can be applied in practice.

ACKNOWLEDGMENT

The authors would like to thank the volunteers who participated in the listening test.

REFERENCES

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

[2] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, Mar. 1998.
[3] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. IEEE Int. Canf. Acoust., Speech, Signal Process. (ICASSP)*, Seattle, WA, May 1998, pp. 285–289.
[4] A. Mouchtaris, J. Van der Spiegel, and P. Mueller, "Non-parallel training for voice conversion by maximum likelihood constrained adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Montreal, QC, Canada, May 2004, pp. 1–4.
[5] ——, "Nonparallel training for voice conversion based on a parameter adaptation approach," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 3, pp. 952–963, May 2006.
[6] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.
[7] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-26, no. 3, pp. 197–210, Jun. 1978.
[8] A. Papoulis, *Probability, Random Variables and Stochastic Processes*. New York: McGraw-Hill, 1991.
[9] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Process.*, vol. 39, no. 4, pp. 795–805, Apr. 1991.
[10] T. V. Sreenivas and P. Kirnapure, "Codebook constrained Wiener filtering for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 383–389, Sept. 1996.
[11] Y. Ephraim and D. Mallah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
[12] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.
[13] Y. Ephraim, D. Mallah, and B.-H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 6, pp. 1846–1856, Dec. 1989.
[14] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 40, no. 2, pp. 725–735, Apr. 1992.
[15] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.
[16] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Signal Process.*, vol. 39, no. 8, pp. 1732–1742, Aug. 1991.
[17] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 373–385, Jul. 1998.
[18] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech Audio Process.*, vol. 5, pp. 497–514, Nov. 1997.
[19] L. Deng, J. Droppo, and A. Acero, "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 568–580, Nov. 2003.
[20] A. Mouchtaris, J. Van der Spiegel, and P. Mueller, "A spectral conversion approach to the iterative Wiener filter for speech enhancement," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Taipei, Taiwan, 2004, pp. 1971–1974.
[21] J. Wu, J. Droppo, L. Deng, and A. Acero, "A noise-robust ASR front-end using Wiener filter constructed from MMSE estimation of clean speech and noise," in *Proc. IEEE Workshop Automatic Speech Recognition Understanding (ASRU)*, 2003, pp. 321–326.
[22] S. Haykin, *Adaptive Filter Theory*. Englewood Cliffs: Prentice-Hall, 1996.
[23] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.
[24] A. Mouchtaris, S. S. Narayanan, and C. Kyriakakis, "Maximum likelihood constrained adaptation for multichannel audio synthesis," in *Conf. Rec. 36th Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, Nov. 2002, vol. I, pp. 227–232.
[25] V. D. Diakoloukas and V. V. Digalakis, "Maximum-likelihood stochastic-transformation adaptation of Hidden Markov Models," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 2, pp. 177–187, Mar. 1999.
[26] A. Kain, "High resolution voice transformation," Ph.D. dissertation, OGI School Sci. Eng., Oregon Health Sci. Univ., Portland, Oct. 2001.
[27] A. Varga and H. J. M. Steeneken, "Assesment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247–251, 1993.
[28] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.
[29] W. B. Kleijn and K. K. Paliwal, Eds., *Speech Coding and Synthesis*. New York: Elsevier, 1995.

Athanasios Mouchtaris (S'02–M'04) received the Diploma degree in electrical engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1997 and the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, in 1999 and 2003, respectively.

From 2003 to 2004 he was a Postdoctoral Researcher in the Electrical and Systems Engineering Department, University of Pennsylvania, Philadelphia. He is currently a Postdoctoral Researcher in the Institute of Computer Science of the Foundation for Research and Technology—Hellas (ICS-FORTH), Heraklion, Crete. He is also a Visiting Professor in the Computer Science Department of the University of Crete, Crete, Greece. His research interests include signal processing for immersive audio environments, spatial audio rendering, multichannel audio modeling, speech synthesis with emphasis on voice conversion, and speech enhancement.

Dr. Mouchtaris is a member of Eta Kappa Nu.

Jan Van der Spiegel (M'72–SM'90–F'02) received the M.S. degree in electromechanical engineering and the Ph.D. degree in electrical engineering from the University of Leuven, Leuven, Belgium, in 1974 and 1979, respectively.

He is currently a Professor of the Electrical and Systems Engineering Department, and the Director of the Center for Sensor Technologies at the University of Pennsylvania, Philadelphia. His primary research interests are in high-speed, low-power analog and mixed-mode VLSI design, biologically-based sensors and sensory information processing systems, microsensor technology, and analog-to-digital converters. He is the author of over 160 journal and conference papers and holds four patents.

Prof. Van der Spiegel is the recipient of the IEEE Third Millennium Medal, the UPS Foundation Distinguished Education Chair, and the Bicentennial Class of 1940 Term Chair. He received the Christian and Mary Lindback Foundation, and the S. Reid Warren Award for Distinguished Teaching, and the Presidential Young Investigator Award. He has served on several IEEE program committees (IEDM, ICCD, ISCAS, and ISSCC) and is currently the technical program Vice-Chair of the International Solid-State Circuit Conference (ISSCC2006). He is an elected member of the IEEE Solid-State Circuits Society and is also the SSCS chapters Chairs coordinator and former Editor of Sensors and Actuators A for North and South America. He is a member of Phi Beta Delta and Tau Beta Pi.

Paul Mueller received the M.D. degree from Bonn University, Bonn, Germany.

He was formerly with the Rockefeller University, New York, and the University of Pennsylvania, Philadelphia, and is currently Chairman of Corticon, Inc, King of Prussia. He has worked on ion channels, lipid bilayers, neural processing of vision and acoustical patterns and VLSI implementation of neural systems.

Panagiotis Tsakalides (M'95) received the Diploma in electrical engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1990 and the Ph.D. degree in electrical engineering from the University of Southern California (USC), Los Angeles, in 1995.

He is an Associate Professor of Computer Science at the University of Crete, Greece, where, from 2004 to 2006, he was the Department Chairman. He is also a Researcher with the Institute of Computer Science, Foundation for Research and Technology-Hellas (FORTH-ICS), Heraklion, Greece. From 1999 to 2002, he was with the Department of Electrical Engineering, University of Patras, Greece. From 1996 to 1998, he was a Research Assistant Professor with the Signal and Image Processing Institute, USC, and he consulted for the U.S. Navy and Air Force. His research interests lie in the field of statistical signal processing with emphasis in non-Gaussian estimation and detection theory and applications in wireless communications, imaging, and multimedia systems. He has coauthored over 60 technical publications in these areas, including 20 journal papers.

Dr. Tsakalides was awarded the IEE's A. H. Reeve Premium in 2002 for the paper (coauthored with P. Reveliotis and C. L. Nikias) "Scalar quantization of heavy tailed signals," published in the October 2000 issue of the *IEE Proceedings—Vision, Image and Signal Processing*.