Single-Sensor Audio Source Separation Using Classification and Estimation Approach and GARCH Modeling

Ari Abramson, *Student Member, IEEE*, and Israel Cohen, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a new algorithm for single-sensor audio source separation of speech and music signals, which is based on generalized autoregressive conditional heteroscedasticity (GARCH) modeling of the speech signals and Gaussian mixture modeling (GMM) of the music signals. The separation of the speech from the music signal is obtained by a simultaneous classification and estimation approach, which enables one to control the tradeoff between residual interference and signal distortion. Experimental results on mixtures of speech and piano music signals have yielded an improved source separation performance compared to using Gaussian mixture models for both signals. The tradeoff between signal distortion and residual interference is controlled by adjusting some cost parameters, which are shown to determine the missed and false detection rates in the proposed classification and estimation approach.**

***Index Terms*—Detection and estimation, generalized autoregressive conditional heteroscedasticity (GARCH), Source separation.**

I. Introduction

SEPARATION of mixed audio signals received by a single microphone has been a challenging problem for many years. Examples of applications include separation of speakers [1], [2], separation of different musical sources (e.g., different musical instruments) [1], [3], [4], separation of speech or singing voice from background music [5]–[8], and signal enhancement in nonstationary noise environments [9]–[13]. In case the signals are received by multiple microphones, spatial filtering may be employed as well as mutual information between the received signals, e.g., see [14] and references therein. However, for the underdetermined case of several sources which are recorded by a single microphone, some *a priori* information is necessary to enable reasonable separation performance. Existing algorithms for single-sensor audio source separation generally deal with two main problems. The first is to obtain appropriate statistical models for the mixed signals, i.e., codebook, and the second problem is the design of a separation algorithm.

Manuscript received September 04, 2007; revised July 31, 2008. Current version published October 17, 2008. This work was supported by the Israel Science Foundation under Grant 1085/05 and by the European Commission under project Memories FP6-IST-035300. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Susanto Rahardja.

The authors are with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Technion City, Haifa 32000, Israel (e-mail: aari@tx.technion.ac.il; icohen@ee.technion.ac.il).

Digital Object Identifier 10.1109/TASL.2008.2005351

In [12] and [13], speech and nonstationary noise signals are assumed to evolve as mixtures of autoregressive (AR) processes in the time domain. The *a priori* statistical information (codebook), which in this case includes the sets of AR prediction coefficients, is obtained by using a training phase. In [3], [4], and [6], the acoustic signals are modeled by Gaussian mixture models (GMMs), and in [9] and [10], the acoustic signals are modeled by hidden Markov models (HMMs) with AR subsources. The trained codebooks provide statistical information about the distinct signals, which enables source separation from signal mixtures. The desired signal may be reconstructed based on the assumed model by minimizing the mean-square error (mse) [4], [6], [10] or by a maximum *a posteriori* (MAP) approach [9]. However, in case of several sources received by a single sensor, separation performances are far from being perfect. Falsely assigning an interfering component to the desired signal may cause an annoying residual interference, while falsely attenuating components of the desired signal may result in signal distortion and perceptual degradation.

GMM and AR-based codebooks are generally insufficient for source separation of statistically rich signals such as speech signals since they only allow a finite set of probability density functions (pdf's) [8], [15]. Recently, generalized autoregressive conditional heteroscedasticity (GARCH) models have been proposed for modeling speech signals for speech enhancement [16]–[19], speech recognition [20], and voice activity detection [21] applications. The GARCH model takes into account the correlation between successive spectral variances and specifies a time-varying conditional variance (volatility) as a function of past spectral variances and squared-absolute values. As a result, the spectral variances may smoothly change along time and the pdf is much less restricted [22]–[24].

In this paper, we propose a novel approach for single-sensor audio source separation of speech and music signals. We consider both problems of codebook design and the ability to control the tradeoff between the residual interference and the distortion of the desired signal. Accordingly, the proposed approach includes a new codebook for speech signals, as well as a new separation algorithm which relies on a simultaneous classification and estimation method. The codebook is based on GARCH modeling of speech signals and Gaussian mixture modeling of music signals. We apply the models to distinctive frequency subbands, and define a specific state for the case that the signal is absent in the observed subband. The proposed separation algorithm relies on integrating a classifier and an estimator while reconstructing each signal. The classifier attempts at classifying

the observed signal into the appropriate hypotheses of each of the signals, and the estimator output is based on the classification. Two methods are proposed for classification and estimation. One is based on simultaneous operations of classification and estimation while minimizing a combined Bayes risk. The second method employs a given (nonoptimal) classifier, and applies an estimator which is optimally designed to yield a controlled level of residual interference and signal distortion. The GARCH model for the speech signal enables smooth covariance matrices with possible state switching. Experimental results demonstrate that for mixtures of speech and piano signals it is more advantageous to model the speech signal by GARCH than GMM, and the codebook generated by the GARCH model yields significantly improved separation performance. In addition, the classification and estimation approach, together with the signal absence state, enables the user to control the tradeoff between distortion of the desired signal caused by missed detection, and amount of residual interference resulting from false detection.

This paper is organized as follows. In Section II, we briefly review codebook-based methods for single-channel audio source separation. We formulate the simultaneous classification and estimation problem for mixtures of signals and derive an optimal solution for the classifier and the combined estimator. Furthermore, we show that a constrained optimization with a given classifier yields the same estimator. In Section III, we define the GARCH codebook which is considered for speech signals and review the recursive conditional variance estimation. In Section IV, we describe the implementation of the proposed algorithm, and in Section V we provide some experimental results for audio separation of speech and music signals.

II. Codebook-Based Separation

Separation of a mixture of signals observed via a single sensor is an ill-posed problem. Some *a priori* information about the mixed signals is generally necessary to enable reasonable reconstructions. Benaroya *et al.* [3]–[5] proposed a GMM for the signal's codebook in the short-time Fourier transform (STFT) domain, and in [9], [10], [12], and [13], mixtures of AR models are considered in the time domain. In each case, a set of clean *similar* signals is used to train the codebooks prior to the separation step. Although the AR processes are defined in the time domain, for process of length N with prediction coefficients $\{1, a_1, \ldots, a_p\}$ and innovation variance σ^2, the covariance matrix is $\sigma^2(A^TA)^{-1}$, where A is an $N \times N$ lower triangular Toeplitz matrix with $[1 \;\; a_1 \;\; \cdots \;\; a_p \;\; 0 \;\; \cdots \;\; 0]^T$ as the first column. If the frame length N tends to infinity, the covariance matrix become circulant and hence diagonalized by the Fourier transform [10], [13]. Accordingly, each set of AR coefficients, together with the excitation variance, corresponds to a specific covariance matrix in the STFT domain similarly to the GMM. Therefore, under any of these models, each framed signal is considered as generated from some specific distribution, which is related to the codebook with some probability, and separation is applied on a frame-by-frame basis.

We now start with brief introduction of existing codebooks and separation algorithms. Let $\mathbf{s}_1, \mathbf{s}_2 \in \mathbb{C}^N$ denote the vectors of the STFT expansion coefficients of signals $s_1(n)$ and $s_2(n)$, respectively, for some specific frame index. Let q_1 and q_2 denote the active states of the codebooks corresponding to signals $\mathbf{s}_1$ and $\mathbf{s}_2$, respectively, with known *a priori* probabilities $p_1(i) \triangleq p(q_1 = i)$, $i = 1, \ldots, m_1$ and $p_2(j) \triangleq p(q_2 = j)$, $j = 1, \ldots, m_2$, and $\sum_i p_1(i) = \sum_j p_2(j) = 1$. Given that $q_1 = i$ and $q_2 = j$, $\mathbf{s}_1$ and $\mathbf{s}_2$ are assumed conditionally zero-mean proper complex-valued Gaussian random vectors (see, e.g., [25], [26, p. 89]) with known diagonal covariance matrices, i.e., $\mathbf{s}_1 \sim \mathcal{CN}(0, \Sigma_1^{(i)})$ and $\mathbf{s}_2 \sim \mathcal{CN}(0, \Sigma_2^{(j)})$.

Based on a given codebook, it is proposed in [4] and [13] to first find the active pair of states $\{i, j\} = \{q_1 = i, q_2 = j\}$ using a MAP criterion

$$\{\hat{i}, \hat{j}\} = \arg\max_{i,j} p(\mathbf{x} \,|\, i, j) p(i, j) \tag{1}$$

where $\mathbf{x} = \mathbf{s}_1 + \mathbf{s}_2$, $p(\cdot \,|\, i, j) = p(\cdot | q_1 = i, q_2 = j)$, and for statistically independent signals $p(i, j) = p_1(i) p_2(j)$. Subsequently, conditioned on these states (i.e., classification), the desired signal may be reconstructed in the minimum mean squared error (mmse) sense by

$$\begin{aligned} \hat{\mathbf{s}}_1 &= E\{\mathbf{s}_1 \,|\, \mathbf{x}, \hat{i}, \hat{j}\} \\ &= \Sigma_1^{(\hat{i})} \left(\Sigma_1^{(\hat{i})} + \Sigma_2^{(\hat{j})}\right)^{-1} \mathbf{x} \\ &\triangleq W_{\hat{i}\hat{j}} \, \mathbf{x} \end{aligned} \tag{2}$$

and similarly[1] $\hat{\mathbf{s}}_2 = W_{\hat{j}\hat{i}} \, \mathbf{x}$. Alternatively [4], [6], [10], the desired signal may be reconstructed in the mmse sense directly from the mixed signal

$$\begin{aligned} \hat{\mathbf{s}}_1 &= E\{\mathbf{s}_1 \,|\, \mathbf{x}\} \\ &= \sum_{i,j} p(i, j | \mathbf{x}) W_{ij} \, \mathbf{x}. \end{aligned} \tag{3}$$

Note that in case of additional uncorrelated stationary noise in the mixed signal, i.e., $\mathbf{x} = \mathbf{s}_1 + \mathbf{s}_2 + \mathbf{d}$ with $\mathbf{d} \sim \mathcal{CN}(0, \Sigma)$, the covariance matrix of the noise signal is added to the covariance matrix of the interfering signal, and then the signal estimators remain in the same forms. Furthermore, without loss of generality, we may restrict ourselves to the problem of restoring the signal $\mathbf{s}_1$ from the observed signal $\mathbf{x}$.

In the following subsections, we introduce two related methods for separation. In Section II-A, we formulate the problem of source separation as a *simultaneous classification and estimation* problem in the sense of statistical decision theory. A classifier is aimed at finding the appropriate states within the codebooks, and the estimator tries to estimate the desired signal based on the given classification. Coupled classifier and estimator jointly minimize a combined Bayes risk, which penalizes for both classification and estimation errors. Relying on the fact that audio signals are generally sparse in the STFT domain, we define additional specific states for the codebook which represent signal absence, and consider false detection of the desired signal and missed detection. The false detection results in under-attenuation of the interfering signal. On the other hand, missed detection of the desired signal may result in removal of desired components and excessive distortion

[1]Note that in this paper the index i always refers to the signal s_1 and the index j refers to the other signal s_2. Therefore, $W_{ji} = \Sigma_2^{(j)}(\Sigma_1^{(i)} + \Sigma_2^{(j)})^{-1}$.

of the separated signals. To allow the user a control over the residual interference and the signal distortion, we introduce cost parameters which are related to missed detection and false detection of the desired signal.

In Section II-B, we introduce a slightly different formulation of optimal estimation under a given classifier. An independent (given) classifier may be applied, for example, by using the MAP classifier (1). Based on this classification, the signal estimation is derived by solving a constrained optimization with respect to the level of residual interference and signal distortion. We denote this approach as *sequential classification and estimation*. We show that in case of degenerated *simultaneous* classification and estimation formulation, closely related solutions can be derived under both approaches.

A. Simultaneous Classification and Estimation

Simultaneous detection and estimation formulation was first proposed by Middleton *et al.* [27], [28]. This scheme assumes coupled operations of detection and estimation which jointly minimize a combined Bayes risk. Recently, a similar approach has been proposed for speech enhancement in nonstationary noise environments [29]. It was shown that applying simultaneous operations of speech detection and estimation in the STFT domain improves the enhanced signal compared to using an estimation only approach. Furthermore, the contribution of the detector is more significant when the interfering signal is highly nonstationary. In this subsection, we develop a simultaneous classification and estimation approach for a codebook-based single-channel audio source separation. By introducing cost parameters for classification errors, the tradeoff between residual interference and signal distortion may be controlled.

Let η denote a classifier for the mixed signal, where η_{ij} indicates that the mixed signal $\mathbf{x}$ is classified to be associated with the pair of states $\{i,j\}$. Let

$$C_{ij}^{\bar{i}\bar{j}}(\mathbf{s},\hat{\mathbf{s}}) \triangleq b_{ij}^{\bar{i}\bar{j}}\|\mathbf{s}-\hat{\mathbf{s}}\|_2^2 \tag{4}$$

denote the combined cost of classification and estimation, where we use a squared-error distortion, and $b_{ij}^{\bar{i}\bar{j}} > 0$ are parameters which impose a penalty for making a decision that $\{i,j\}$ is the active pair while actually $\mathbf{s}_1$ was generated with covariance matrix $\Sigma_1^{(\bar{i})}$ and $\mathbf{s}_2$ with covariance matrix $\Sigma_2^{(\bar{j})}$ (i.e., $q_1=\bar{i}$ and $q_2=\bar{j}$). The combined risk of classification and estimation is then given by

$$R=\sum_{i,j}\sum_{\bar{i},\bar{j}}\int\int C_{ij}^{\bar{i}\bar{j}}(\mathbf{s}_1,\hat{\mathbf{s}}_1)p(\mathbf{x}\,|\,\mathbf{s}_1,\bar{i},\bar{j}) \times p(\mathbf{s}_1\,|\,\bar{i},\bar{j})p(\bar{i},\bar{j})p(\eta_{ij}\,|\,\mathbf{x})d\mathbf{s}_1\,d\mathbf{x}. \tag{5}$$

The simultaneous classification and estimation is aimed at finding the optimal estimator and classifier which jointly minimize the combined risk

$$\min_{\eta_{ij},\hat{\mathbf{s}}_1}\{R\}. \tag{6}$$

To derive a solution to (6), we first note that the signal $\mathbf{s}_1$ is independent of the value of q_2, hence $p(\mathbf{s}_1\,|\,\bar{i},\bar{j}) = p(\mathbf{s}_1\,|\,\bar{i})$. Similarly, $\mathbf{x}$ given $\mathbf{s}_1$ is independent of the value of q_1. Accordingly, $r_{ij}^{\bar{i}\bar{j}}(\mathbf{x},\hat{\mathbf{s}}_1)$ which is defined by

$$r_{ij}^{\bar{i}\bar{j}}(\mathbf{x},\hat{\mathbf{s}}_1)=\int C_{ij}^{\bar{i}\bar{j}}(\mathbf{s}_1,\hat{\mathbf{s}}_1)p(\mathbf{x}\,|\,\mathbf{s}_1,\bar{j})p(\mathbf{s}_1\,|\,\bar{i})d\mathbf{s}_1 \tag{7}$$

denotes the average risk related to a decision η_{ij} when the true pair is $\{\bar{i},\bar{j}\}$. The combined risk (5) can be written as

$$R=\sum_{i,j}\int p(\eta_{ij}\,|\,\mathbf{x})\sum_{\bar{i},\bar{j}}p(\bar{i},\bar{j})r_{ij}^{\bar{i}\bar{j}}(\mathbf{x},\hat{\mathbf{s}}_1)d\mathbf{x}. \tag{8}$$

The classifier's decision for a given observation is nonrandom. Therefore, given the observed signal $\mathbf{x}$, the optimal estimator under a decision η_{ij} made by the classifier [i.e., $p(\eta_{ij}\,|\,\mathbf{x})=1$ for a particular pair $\{i,j\}$] is obtained by

$$\hat{\mathbf{s}}_{1,ij}=\arg\min_{\hat{\mathbf{s}}_1}\sum_{\bar{i},\bar{j}}p(\bar{i},\bar{j})r_{ij}^{\bar{i}\bar{j}}(\mathbf{x},\hat{\mathbf{s}}_1). \tag{9}$$

Substituting (7) into (9) and setting the derivative to be equal to zero, we obtain the optimal estimate under η_{ij}

$$\hat{\mathbf{s}}_{1,ij}=\frac{\sum_{\bar{i}\bar{j}}b_{ij}^{\bar{i}\bar{j}}\,p(\mathbf{x}\,|\,\bar{i},\bar{j})p(\bar{i},\bar{j})W_{\bar{i}\bar{j}}\,\mathbf{x}}{\sum_{\bar{i}\bar{j}}b_{ij}^{\bar{i}\bar{j}}\,p(\mathbf{x}\,|\,\bar{i},\bar{j})p(\bar{i},\bar{j})} \triangleq G_{ij}\,\mathbf{x}. \tag{10}$$

The derivation of (10) is given in Appendix A. Note that in case the parameters $b_{ij}^{\bar{i}\bar{j}}$ are all equal, then the estimator (10) reduces to the mmse estimator (3) and the estimation does not depend on the classification rule.

The average risk $r_{ij}^{\bar{i}\bar{j}}(\mathbf{x},\hat{\mathbf{s}}_1)$ is a function of the observed mixed signal and the optimal estimate under η_{ij}. Let $\mathbf{1}$ denote a column vector of ones. Then, by substituting (10) into (7), we obtain (see Appendix B)

$$r_{ij}^{\bar{i}\bar{j}}(\mathbf{x},\hat{\mathbf{s}}_1)=b_{ij}^{\bar{i}\bar{j}}p(\mathbf{x}|\bar{i},\bar{j}) \times\left[\mathbf{x}^H\left(W_{\bar{i}\bar{j}}^2-2W_{\bar{i}\bar{j}}G_{ij}\right)\mathbf{x}+\mathbf{1}^T\Sigma_2^{(\bar{j})}W_{\bar{i}\bar{j}}\mathbf{1}\right]. \tag{11}$$

From (6) and (8), the optimal classification rule $\eta_{ij}(\mathbf{x})$ is obtained by minimizing the weighted average risks over all pairs of states

$$\min_{ij}\sum_{\bar{i},\bar{j}}p(\bar{i},\bar{j})r_{ij}^{\bar{i}\bar{j}}(\mathbf{x},\hat{\mathbf{s}}_1). \tag{12}$$

If we consider the degenerated case of equal parameters $b_{ij}^{\bar{i}\bar{j}}$, then the averaged risk $r_{ij}^{\bar{i}\bar{j}}(\mathbf{x},\hat{\mathbf{s}}_1)$ does not depend on i,j, and therefore there is no specific pair which minimizes (12). However, as already mentioned above, in this case there is no need for a classification since the estimator does not depend on the decision rule.

To summarize, minimizing the combined Bayes risk is obtained by first evaluating the optimal gain matrix G_{ij} under

each pair $\{i,j\}$ using (10), and subsequently the optimal classifier chooses the appropriate pair (and the appropriate gain matrix) using (11) and (12). The combined solution guaranties minimum combined Bayes risk [28].

The selection of the parameters $b_{ij}^{\bar{i}\bar{j}}$ is application dependent, since these parameters determine the penalty for choosing each set of states compared to all other sets. Recall we would like to define specific states for signal absence, we consider from now on that the signal states are $q_1 \in \{0,1,\ldots,m_1\}$ and $q_2 \in \{0,1,\ldots,m_2\}$ where $q_1 = 0$ and $q_2 = 0$ are the signal absence states. Accordingly, we define separable parameters $b_{ij}^{\bar{i}\bar{j}} = b_i^{\bar{i}} b_j^{\bar{j}}$, where b_i^0 with $i \neq 0$ is related to the cost of false detection and $b_0^{\bar{i}}$ with $\bar{i} \neq 0$ is related to the cost of missed detection of the desired signal. Specifically, we define

$$b_i^{\bar{i}} = \begin{cases} b_{1,\mathrm{m}} & i = 0,\ \bar{i} \neq 0 \\ b_{1,\mathrm{f}} & i \neq 0,\ \bar{i} = 0 \\ 1 & \text{o.w.} \end{cases} \tag{13}$$

with $b_{1,\mathrm{m}},\ b_{1,\mathrm{f}} > 0$, and for signal s_2, $b_j^{\bar{j}}$ is defined similarly (with parameters $b_{2,\mathrm{m}}$ and $b_{2,\mathrm{f}}$). By using this definition, we practically assume equal parameters (i.e., one) for all cases except for missed detection and false detection. As can be seen from (11) and (12), higher $b_{2,\mathrm{m}}$ (or $b_{2,\mathrm{f}}$) results in larger average risk which corresponds to this decision, and therefore, lower chances for the optimal detector to take this decision. However, as can be seen from (10), the high-valued parameter raises the contribution of the corresponding state on the system estimate. If a parameter is smaller than one, then the chances of the detector to take this decision are higher, but, as the estimator (10) compensates for wrong decisions, this contribution on the system estimate would be low. Missing to detect the desired signal results, in general, in removal of desired signal components and therefore distorts the desired signal estimate. On the other hand, false detection may result in residual interference. By affecting both the decision rule and the corresponding estimation, these parameters help to control the tradeoff between residual interference when the desired signal is absent (resulting from false detection) and the distortion of the desired signal caused by missed detection.

The computational complexity of the simultaneous classification and estimation approach is higher than that associated with the sequential MAP classification and mmse estimation (1)–(2), or the mmse estimator (3). However, the estimator (10) is optimal, not only when combined with the optimal classifier (12), but also when combined with any given classifier [28]. Therefore, this estimator may be combined with a suboptimal classifier [e.g., the MAP classifier given by (1)] to reduce the computational requirements, while still using parameters which compensate for false classification. In the following subsection, we discuss this option of employing a nonideal classifier and show that the same estimator (10) can be obtained by solving a constrained optimization problem. In this problem formulation, it is shown that the cost parameters may also have the interpretation of Lagrange multipliers.

B. Sequential Classification and Estimation

The application of a given classifier (e.g., a MAP classifier) followed by an estimator is shown in Fig. 1. We denote this

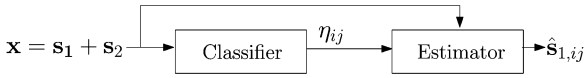

Fig. 1. Cascade classification and estimation scheme.

scheme as *sequential classification and estimation*. In order to simplify the derivation, we assume in this subsection only signal absence or presence states $i,j \in \{0,1\}$ (i.e., $m_1 = m_2 = 1$), where $i = 0$ and $i = 1$ represent presence and, respectively, absence of $\mathbf{s}_1$, and j similarly specifies the state of $\mathbf{s}_2$. The classifier is generally not ideal and may suffer from miss and false detections. Therefore, under false decision that the signal is absent when actually the signal is present, we may want to control the distortion level, while under false detection of signal components we wish to control the level of residual interference. Under the two hypotheses, the mean signal distortion is defined by

$$\bar{\varepsilon}_d^2(\mathbf{x}) \triangleq p(q_1 = 1|\mathbf{x}) E\left\{ \|\mathbf{s}_1 - \hat{\mathbf{s}}_1\|_2^2 \mid q_1 = 1, \mathbf{x} \right\} \tag{14}$$

and the mean residual interference is defined by

$$\bar{\varepsilon}_r^2(\mathbf{x}) \triangleq p(q_1 = 0 \mid \mathbf{x}) E\left\{ \|\mathbf{s}_1 - \hat{\mathbf{s}}_1\|_2^2 \mid q_1 = 0, \mathbf{x} \right\}. \tag{15}$$

Therefore, for a decision that signal is absent (i.e., η_{0j}) we have the following problem:

$$\hat{\mathbf{s}}_{1,0j} = \arg\min_{\hat{\mathbf{s}}_1} p(q_1 = 0|\mathbf{x}) E\left\{ \|\mathbf{s}_1 - \hat{\mathbf{s}}_1\|_2^2 \mid q_1 = 0, \mathbf{x} \right\} \quad \text{s.t. } \bar{\varepsilon}_d^2(\mathbf{x}) \leq \sigma_d^2, \tag{16}$$

while for a signal-presence decision (η_{1j}) we have

$$\hat{\mathbf{s}}_{1,1j} = \arg\min_{\hat{\mathbf{s}}_1} p(q_1 = 1 \mid \mathbf{x}) E\left\{ \|\mathbf{s}_1 - \hat{\mathbf{s}}_1\|_2^2 \mid q_1 = 1, \mathbf{x} \right\} \quad \text{s.t.} \quad \bar{\varepsilon}_r^2(\mathbf{x}) \leq \sigma_r^2 \tag{17}$$

where σ_d^2 and σ_r^2 are bounds for the mean distortion and mean residual interference, respectively. The optimal estimator can be obtained by using a method similar to [30] and [31]. Under η_{0j} the Lagrangian is defined by (e.g., [32])

$$L_d(\hat{\mathbf{s}}_1, \mu_d) = p(q_1 = 0 \mid \mathbf{x}) E\left\{ \|\mathbf{s}_1 - \hat{\mathbf{s}}_1\|_2^2 \mid q_1 = 0, \mathbf{x} \right\} + \mu_d \left(\bar{\varepsilon}_d^2(\mathbf{x}) - \sigma_d^2 \right) \tag{18}$$

and

$$\mu_d(\bar{\varepsilon}_d^2(\mathbf{x}) - \sigma_d^2) = 0 \quad \text{for} \quad \mu_d \geq 0. \tag{19}$$

Under η_{1j}, the Lagrangian $L_r(\hat{\mathbf{s}}_1, \mu_r)$ is defined similarly using μ_r and $\bar{\varepsilon}_r^2(\mathbf{x})$. Then, $\hat{\mathbf{s}}_{1,0j}$ (or $\hat{\mathbf{s}}_{1,1j}$) is a stationary feasible point if it satisfies the gradient equation of the appropriate Lagrangian [i.e., $L_d(\hat{\mathbf{s}}_1, \mu_d)$ or $L_r(\hat{\mathbf{s}}_1, \mu_r)$]. From $\bigtriangledown_{\hat{\mathbf{s}}_1} L_d(\hat{\mathbf{s}}_1, \mu_d) = 0$ we have[2] (20), as shown at the bottom of the next page, where

$$\tilde{\mu}_d^{\bar{i}} = \begin{cases} \mu_d & \bar{i} = 1 \\ 1 & \bar{i} = 0 \end{cases}. \tag{21}$$

[2]Note that as shown in [30] and [31], there is no closed form solution for the value of the Lagrange multiplier. Instead it is used as a non-negative parameter.

Similarly, under signal-presence decision we have

$$\hat{\mathbf{s}}_{1,1j} = \frac{\sum_{\bar{i}\bar{j}} \tilde{\mu}_r^{\bar{i}}\, p(\bar{i},\bar{j}\,|\,\mathbf{x}) W_{\bar{i}\bar{j}}\mathbf{x}}{\sum_{\bar{i}\bar{j}} \tilde{\mu}_r^{\bar{i}}\, p(\bar{i},\bar{j}\,|\,\mathbf{x})} \tag{22}$$

with

$$\tilde{\mu}_r^{\bar{i}} = \begin{cases} \mu_r & \bar{i}=0 \\ 1 & \bar{i}=1 \end{cases}. \tag{23}$$

Therefore, in general we can write

$$\hat{\mathbf{s}}_{1,ij} = \frac{\sum_{\bar{i}\bar{j}} \mu_i^{\bar{i}} p(\bar{i},\bar{j}\,|\,\mathbf{x}) W_{\bar{i}\bar{j}}\mathbf{x}}{\sum_{\bar{i}\bar{j}} \mu_i^{\bar{i}} p(\bar{i},\bar{j}\,|\,\mathbf{x})} \tag{24}$$

with

$$\mu_i^{\bar{i}} = \begin{cases} \mu_d & i=0, \bar{i}=1 \\ \mu_r & i=1, \bar{i}=0 \\ 1 & \text{o.w.} \end{cases} \tag{25}$$

and the estimator (24) is the same as (10) with $b_j^{\bar{j}} = 1$, $b_{1,\mathrm{m}} = \mu_d$, and $b_{1,\mathrm{f}} = \mu_r$ in (13). Therefore, we can identify the parameters $b_{ij}^{\bar{i}\bar{j}}$ as non-negative Lagrange multipliers of a constrained optimization problem. In addition, if $b_{1,\mathrm{m}}$ (or $b_{1,\mathrm{f}}$) equals zero, then the corresponding Lagrange multiplier also reduces to zero and the constraint in (16) [or (17)] is inapplicable. Therefore, the problem reduces to a standard conditional mmse problem, which results in the estimator (2) which assumes a perfect classifier.

The main difference between the problem formulations in Sections II-A and II-B is that the former defines a classifier and a coupled estimator which are designed to minimize a combined Bayes risk, while the latter assumes a given classifier, and formulates a constrained optimization problem in order to find the optimal estimator for the given classification rule.

III. GMM Versus GARCH Codebook

In this section, we introduce a new codebook for mixtures of speech and music signals. GMM was used in [4]–[6] for generating codebooks for speech signals as well as for music signals in the STFT domain, under the assumption of diagonal covariance matrices. The covariance matrices and the *a priori* state probabilities are estimated by either maximizing the log-likelihood of the trained signal using expectation-maximization algorithm [9], [33], or by using the k-means vector quantization algorithm [9], [34]. Using a finite-state model with predetermined densities as in the case of GMM, mixture of AR models or HMM with AR subsources, the diagonal vector of the covariance matrices can take values only from a specific subspace of $\mathbb{R}_+^N$ spanned by the given codewords. This limitation for the pdf's may restrict the usage of these models for statistically rich signals such as speech [8].

GARCH is a statistical model which explicitly parameterizes a time-varying conditional variance using past variances and squared absolute values, while considering volatility clustering and excess kurtosis (i.e., heavy-tailed distribution) [22]. Expansion coefficients of speech signals in the STFT domain, are clustered in the sense that successive magnitudes at a fixed frequency bin are highly correlated [24]. GARCH model has been found useful for modeling speech signals in speech enhancement applications [16], [18], [19], speech recognition [20], and voice activity detection [21]. It has been shown [18] that spectral variance estimation resulting from this model is a generalization of the decision-directed estimator [35] with improved tracking of the speech spectral volatility. Therefore, we propose in this paper to use GMM for modeling the music signal (say s_2) and GARCH model with several states for the speech signal (say s_1).

According to the GMM formalism, $p_2(j)$ is the *a priori* probability for the active state $q_2 = j$, where conditioning on $q_2 = j$, the vector in the STFT domain $\mathbf{s}_2 \sim \mathcal{CN}(0, \Sigma_2^{(j)})$. For defining the GARCH modeling we first let $\mathbf{s}_1(\ell)$ denote the ℓth frame of s_1 in the STFT domain. We assume that $\mathbf{s}_1(\ell)$ is a mixture of GARCH processes of order (1,1). Then, given that $q_1(\ell) = i_\ell$ is the active state at frame ℓ, $\mathbf{s}_1(\ell)$ has a complex-normal pdf with zero mean and a diagonal covariance matrix $\Sigma_1^{(i_\ell)} = \mathrm{diag}\,\{\boldsymbol{\lambda}_{\ell|\ell-1}^{(i_\ell)}\}$. The *conditional variance* vector $\boldsymbol{\lambda}_{(\ell|\ell-1}^{i_\ell)}$ is the vector of variances at frame ℓ conditioning on the information up to frame $\ell-1$. This conditional variance is a linear function of the previous conditional variance and squared absolute value

$$\boldsymbol{\lambda}_{\ell|\ell-1}^{(i_\ell)} = \lambda_{\min}^{(i_\ell)}\mathbf{1} + \alpha^{(i_\ell)}\mathbf{s}_1(\ell-1)\odot\mathbf{s}_1^*(\ell-1) + \beta^{(i_\ell)}\left(\boldsymbol{\lambda}_{\ell-1|\ell-2}^{(i_{\ell-1})} - \lambda_{\min}^{(i_{\ell-1})}\mathbf{1}\right) \tag{26}$$

$$\begin{aligned}
\hat{\mathbf{s}}_{1,0j} &= \frac{p(q_1=0\,|\,\mathbf{x})E\{\mathbf{s}\,|\,q_1=0,\mathbf{x}\} + \mu_d\, p(q_1=1\,|\,\mathbf{x})E\{\mathbf{s}\,|\,q_1=1,\mathbf{x}\}}{p(q_1=0\,|\,\mathbf{x}) + \mu_d\, p(q_1=1\,|\,\mathbf{x})} \\
&= \frac{p(q_1=0\,|\,\mathbf{x})\sum_{\bar{j}} p(\bar{j}\,|\,q_1=0,\mathbf{x})E\{\mathbf{s}\,|\,q_1=0,\bar{j},\mathbf{x}\}}{\sum_{\bar{j}} p(q_1=0,\bar{j}\,|\,\mathbf{x}) + \mu_d\sum_{\bar{j}} p(q_1=1,\bar{j}\,|\,\mathbf{x})} \\
&\quad + \frac{\mu_d p(q_1=1|\,\mathbf{x})\sum_{\bar{j}} p(\bar{j}\,|\,q_1=1,\mathbf{x})E\{\mathbf{s}\,|\,q_1=1,\bar{j},\mathbf{x}\}}{\sum_{\bar{j}} p(q_1=0,\bar{j}\,|\,\mathbf{x}) + \mu_d\sum_{\bar{j}} p(q_1=1,\bar{j}\,|\,\mathbf{x})} \\
&= \frac{\sum_{\bar{j}} p(q_1=0,\bar{j}\,|\,\mathbf{x})W_{0\bar{j}}\mathbf{x} + \mu_d\sum_{\bar{j}} p(q_1=1,\bar{j}\,|\,\mathbf{x})W_{1\bar{j}}\mathbf{x}}{\sum_{\bar{j}} p(q_1=0,\bar{j}\,|\,\mathbf{x}) + \mu_d\sum_{\bar{j}} p(q_1=1,\bar{j}\,|\,\mathbf{x})} \\
&= \frac{\sum_{\bar{i}\bar{j}} \tilde{\mu}_d^{\bar{i}}\, p(\bar{i},\bar{j}\,|\,\mathbf{x})W_{\bar{i}\bar{j}}\mathbf{x}}{\sum_{\bar{i}\bar{j}} \tilde{\mu}_d^{\bar{i}}\, p(\bar{i},\bar{j}\,|\,\mathbf{x})}
\end{aligned} \tag{20}$$

where $\odot$ denotes a term-by-term vector multiplication, * denotes complex conjugate, and $\lambda_{\min}^{(i_\ell)} > 0$ and $\alpha^{(i_\ell)}, \beta^{(i_\ell)} \geq 0$ for $i_\ell = 0, 1, \ldots, m_1$ are sufficient conditions for the positivity of the conditional variance [18], [19]. In addition, $\alpha^{(i_\ell)} + \beta^{(i_\ell)} < 1$ for all i_ℓ is a sufficient condition for a finite unconditional variance[3] [23]. The conditional density resulting from (26) is time varying and depends on all past values (through previous conditional variances) and also on the regime path up to the current time. While $\lambda_{\min}^{(i)}$ set the lower bounds for the conditional variances in each state, the parameters $\alpha^{(i)}$ and $\beta^{(i)}$ set the volatility level and the autoregression behavior of the conditional variances. Note that this model is a degenerated case of the Markov-switching GARCH (MS-GARCH) model [18], [37], [38]. In the MS-GARCH model, the sequence of states is a first-order Markov chain with state transition probabilities $p(q_1(\ell) = i_\ell \,|\, q_1(\ell-1) = i_{\ell-1})$. However, to reduce the model complexity and to allow a simpler online estimation procedure under the presence of a highly nonstationary interfering signal, we assume here that the state transition probabilities equal the *a priori* state probabilities, i.e., $p(q_1(\ell) = i_\ell \,|\, q_1(\ell-1) = i_{\ell-1}) = p_1(i_\ell)$, similarly to the assumption used in [4] for the GMM approach.

It can be seen from (26) that the vector of conditional variances $\boldsymbol{\lambda}_{\ell|\ell-1}^{(i_\ell)}$ may take any values in $\mathbb{R}_+^N$ with lower bound $\lambda_{\min}^{(i_\ell)}$ for each entry. However, even if the active state is known, the covariance matrix $\Sigma_1^{(i_\ell)}$ (or the vector of conditional variances $\boldsymbol{\lambda}_{\ell|\ell-1}^{(i_\ell)}$) is unknown and should be reconstructed recursively using all previous signal values and active states. Moreover, since both $\mathbf{s}_1$ and the Markov chain are random processes, the vector of conditional variances is also a random process which follows (26). As we only have a mixed observation, we may estimate this random process of conditional variances based on the recursive estimation algorithm proposed in [18]. Assume that we have an estimate for the set of conditional variances at frame ℓ based on information up to frame $\ell - 1$, $\hat{\Lambda}_\ell \triangleq \{\hat{\boldsymbol{\lambda}}_{\ell|\ell-1}^{(i_\ell)}\}_{i_\ell}$, then, following the model definition an mmse estimate of the next-frame conditional variance follows:

$$\begin{aligned}\hat{\boldsymbol{\lambda}}_{\ell+1|\ell}^{(i_{\ell+1})} &= E\left\{\boldsymbol{\lambda}_{\ell+1|\ell}^{(q_1(\ell+1))} \,|\, q_1(\ell+1) = i_{\ell+1}, \hat{\Lambda}_\ell, \mathbf{x}(\ell)\right\} \\ &= \lambda_{\min}^{(i_{\ell+1})}\mathbf{1} + \alpha^{(i_{\ell+1})} E\{\mathbf{s}_1(\ell) \odot \mathbf{s}_1^*(\ell) \,|\, \hat{\Lambda}_\ell, \mathbf{x}(\ell)\} \\ &\quad + \beta^{(i_{\ell+1})} E\left\{\boldsymbol{\lambda}_{\ell|\ell-1}^{(q_1(\ell))} \,|\, \hat{\Lambda}_\ell, \mathbf{x}(\ell)\right\} \\ &\quad - \beta^{(i_{\ell+1})} E\left\{\lambda_{\min}^{(q_1(\ell))} \,|\, \hat{\Lambda}_\ell, \mathbf{x}(\ell)\right\}\mathbf{1} \end{aligned} \tag{27}$$

for $i_{\ell+1} = 0, 1, \ldots, m_1$. Using

$$\begin{aligned}\hat{\boldsymbol{\lambda}}_{\ell|\ell}^{(i_\ell, j_\ell)} &\triangleq E\{\mathbf{s}_1(\ell) \odot \mathbf{s}_1^*(\ell) \,|\, i_\ell, j_\ell, \hat{\Lambda}_\ell, \mathbf{x}(\ell)\} \\ &= \hat{\Sigma}_1^{(i_\ell)}\left(\hat{\Sigma}_1^{(i_\ell)} + \Sigma_2^{(j_\ell)}\right)^{-1}\Big[\Sigma_2^{(j_\ell)}\mathbf{1} \\ &\quad + \hat{\Sigma}_1^{(i_\ell)}\left(\hat{\Sigma}_1^{(i_\ell)} + \Sigma_2^{(j_\ell)}\right)^{-1}\mathbf{x}(\ell) \odot \mathbf{x}^*(\ell)\Big] \end{aligned} \tag{28}$$

[3]For a necessary and sufficient condition, see [36].

we obtain

$$\begin{aligned} &E\{\mathbf{s}_1(\ell) \odot \mathbf{s}_1^*(\ell) \,|\, \hat{\Lambda}_\ell, \mathbf{x}(\ell)\} \\ &\quad = \sum_{i_\ell, j_\ell} p(i_\ell, j_\ell \,|\, \hat{\Lambda}_\ell, \mathbf{x}(\ell)) \\ &\qquad \times E\{\mathbf{s}_1(\ell) \odot \mathbf{s}_1^*(\ell) \,|\, i_\ell, j_\ell, \hat{\Lambda}_\ell, \mathbf{x}(\ell)\} \\ &\quad = \sum_{i_\ell, j_\ell} p(i_\ell, j_\ell \,|\, \hat{\Lambda}_\ell, \mathbf{x}(\ell))\hat{\boldsymbol{\lambda}}_{\ell|\ell}^{(i_\ell, j_\ell)} \end{aligned} \tag{29}$$

$$\begin{aligned} &E\left\{\boldsymbol{\lambda}_{\ell-1|\ell-2}^{(q_1(\ell))} \,|\, \hat{\Lambda}_\ell, \mathbf{x}(\ell)\right\} \\ &\quad = \sum_{i_{\ell-1}, j_{\ell-1}} p(i_{\ell-1}, j_{\ell-1} \,|\, \hat{\Lambda}_\ell) E\left\{\boldsymbol{\lambda}_{\ell-1|\ell-2}^{(q_1(\ell))} \,|\, i_\ell, \hat{\Lambda}_\ell, \mathbf{x}(\ell)\right\} \\ &\quad \approx \sum_{i_{\ell-1}, j_{\ell-1}} p(i_{\ell-1}, j_{\ell-1} \,|\, \hat{\Lambda}_\ell)\hat{\boldsymbol{\lambda}}_{\ell-1|\ell-2}^{(i_{\ell-1})} \end{aligned} \tag{30}$$

and

$$\begin{aligned} &E\left\{\lambda_{\min}^{(q_1(\ell))} \,|\, \hat{\Lambda}_\ell, \mathbf{x}(\ell)\right\} \\ &\quad = \sum_{i_{\ell-1}, j_{\ell-1}} p(i_{\ell-1}, j_{\ell-1} \,|\, \hat{\Lambda}_\ell)\lambda_{\min}^{(i_{\ell-1})}. \end{aligned} \tag{31}$$

A detailed recursive estimation algorithm is given in [18]. The model parameters, i.e., $\{\lambda_{\min}^{(i_\ell)}, \alpha^{(i_\ell)}, \beta^{(i_\ell)}\}_{i_\ell=0}^{m_1}$, can be estimated from a training set using a maximum-likelihood approach [18], [23], [37] or may be evaluated as proposed in [19] such that each state would represent a different level of the optional dynamic range of the signal's energy. By using the recursive estimation algorithm, we evaluate for each time frame ℓ and for each state i_ℓ an estimate of the spectral covariance matrix $\hat{\Sigma}_1^{(i_\ell)}$ which is required for the separation algorithm. Hence, the sets $\{\hat{\Sigma}_1^{(i_\ell)}\}_{i_\ell}$ and $\{p_1(i_\ell)\}_{i_\ell}$ together with the GMM for the background music signal may be employed by the classification and estimation procedure to obtain an estimate for each signal. Note that for the GMM, each state defines a specific pdf which is known *a priori*, while for the mixing GARCH model the covariance matrices in each state are time-varying and are recursively reconstructed.

IV. Implementation of the Algorithm

The existing GMM-, AR-, and HMM-based algorithms, generally estimate each frame of the signal in the STFT domain using a vector formulation. However, many spectral enhancement algorithms for speech signals treat each frequency bin separately, e.g., [35], [39], and [40]. The application of subband-based audio processing algorithms have been proposed for automatic speech recognition, e.g., [41], [42], speech enhancement [19], and also for single-channel source separation [15]. Instead of applying a statistical model for the whole frame, each subband is assumed to follow a different statistical model. Considering the GARCH modeling, the parameters $\lambda_{\min}^{(i)}$ specify the lower bounds for the conditional variances under each state. Since speech signals are generally characterized by lower energy levels in higher frequency bands, it is advantageous to apply different model parameters in different subbands, as proposed in [19].

For the implementation of the proposed algorithm, we assume $K < N$ linearly spaced frequency subbands for each frame with

independent model parameters. Moreover, the sparsity of the expansion coefficients in the STFT domain (of both speech and music signals) implies that in a specific time-frame the signal may be present in some of the frequency subbands and absent (or of negligible energy) in others. Therefore, we define a specific state for signal absence in each subband $k \in \{1, \ldots, K\}$, $q_1 = 0$ and $q_2 = 0$. For these states, the pdf is assumed to be a zero-mean complex Gaussian with $\sigma^2_{\min,k} I$ covariance matrix. Note that in the GMM case, each state corresponds to a specific predetermined Gaussian density while in the GARCH case, by setting $\alpha^{(0)} = \beta^{(0)} = 0$ and $\lambda^{(0)}_{\min} = \sigma^2_{\min,k}$ for the kth subband, the covariance under $q_1 = 0$ is also time invariant and equals $\sigma^2_{\min,k} I$. In our experimental study, independent models are assumed for each subband, and therefore the model training and both the conditional variance estimation (in case of speech signal) and the separation algorithm are applied independently in each subband. However, in general, some overlap may be considered between adjacent subbands to allow some dependency between them, as well as cross-band state probabilities, as proposed in [15].

Prior to source separation, both the GMM and GARCH models need to be estimated using a set of training signals. In case of GMM, for each state $j \neq 0$ we need to estimate (for each subband independently) the diagonal vector of the covariance matrix $\Sigma_2^{(j)}$, and the state probability $p_2(j)$. For the GARCH modeling, the state probabilities are also required; however, only three scalars are needed to represent the covariance matrix for any $i \neq 0$: $\lambda^{(i)}_{\min}$, $\alpha^{(i)}$, and $\beta^{(i)}$. In our application, the GMM is trained by using the k-means vector quantization algorithm [9], [34]. This model is sensitive to the similarity between the training signals and the desired signal in the mixture, and to achieve good representation, the spectral shapes in the trained and mixed signals need to be closely related, as applied, e.g., in [4], [5], and [13]. For training the GARCH model, we use the method proposed in [19]. Accordingly, the training signals are used only to calculate the peak energy in each of the subbands. Then, we set $\lambda^{(0)}_{\min} = \sigma^2_{\min,k}$, and $\alpha^{(0)} = \beta^{(0)} = 0$. For the speech presence states, $i \in \{1, \ldots, m_1\}$, the parameters are chosen as follows. The lower bounds $\lambda^{(i)}_{\min}$ are log-spaced in the dynamic range, i.e., between $\lambda^{(0)}_{\min}$ and the peak energy. The parameters $\beta^{(i)}$ are experimentally set to 0.8 and $\alpha^{(i)}$ are evaluated for each subband independently such that the stationary variance in the subband, under an immutable state, would be equal to the lower bound for the next state (see [19] for details). This approach yields reasonable results since it enables to represent the whole dynamic range of the signals energy while the conditional variance is updated each frame by using past observation and past conditional variance. In addition, since only the peak energy is required for each subband, this approach has relatively low sensitivity to the training set, and only the peak energy levels need to be similar to that of the test set.

A block diagram of the proposed separation algorithm is shown in Fig. 2 when considering a single band (in practice, a similar process is applied in each subband independently). The observed signal is first transformed into the STFT domain. Then, two steps are applied for each frame ℓ. First, the GARCH

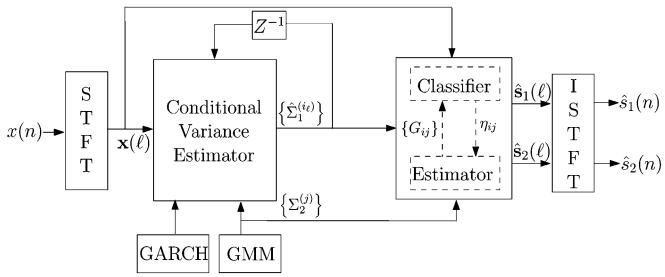

Fig. 2. Block diagram of the proposed algorithm.

conditional covariance matrices $\{\hat{\Sigma}_1^{(i_\ell)} = \mathrm{diag}(\hat{\boldsymbol{\lambda}}^{(i_\ell)}_{\ell|\ell-1})\}_{i_\ell}$ are updated using (28) for any pair $\{i_\ell, j_\ell\}$, and then propagated one frame ahead using (27) to yield the conditional variance estimate for the next frame. Second, using the sets $\{\hat{\Sigma}_1^{(i_\ell)}\}$ and $\{\Sigma_2^{(j)}\}$ the simultaneous classification and estimation method is applied yielding each of the estimates $\hat{\mathbf{s}}_1(\ell)$ and $\hat{\mathbf{s}}_2(\ell)$. Finally, the desired signals are obtained by inverse transforming the signal into the time domain.

Considering a simultaneous classification and estimation approach, as proposed in Section II-A, the interrelations between the classifier and the estimator are employed such that the classification rule is calculated by using the set of gain matrices $\{G_{ij}\}$, and the classifier's output η_{ij} specifies the gain matrix to be used. However, a cascade of classification and estimation (as considered in Section II-B) may be applied as the classification and estimation block to enable a suboptimal solution with lower computational cost. In fact, the computational complexity of this suboptimal method is comparable to that of the mmse estimator (3) since the *a posteriori* probabilities required for the MAP classifier are used also in the estimation step.

V. Experimental Results

In this section, we present experimental results for evaluating the performance of the proposed algorithm. In Section V-A, we describe the experimental setup in our evaluation, and the objective quality measures. Then, in Section V-B, we present experimental results. The experimental results are focused on 1) evaluating the performance of the proposed codebook compared to using a GMM-only model (while using mmse estimation for both codebooks), and 2) evaluating the performance of the proposed simultaneous classification and estimation approach in the sense of signal distortion and residual interference.

A. Experimental Setup and Quality Measures

In our experimental study, we consider speech signals mixed with piano music of about the same level. In the test set of our experimental study, speech signals are taken from the TIMIT database [43] and include eight different utterances by eight different speakers, half male and half female. The speech signals are mixed with two different piano compositions (*Für Elise* by L. van Beethoven and *Romance O' Blue* by W. A. Mozart) to yield 16 different mixed signals. For each of the piano signals, the first 10 s are used to create the mixing signals while the rest of each composition (about 4 min each) is used for training the

model. For training the models to speech signals, a set of signals which are not on the test set was used, with half male and half female (about 30 s length). All signals in the experiment are normalized to the same energy level, and sampled at 16 kHz and transformed into the STFT domain using half-overlapping Hamming window of 32-ms length. The GMM parameters (for the piano model) are trained using the k-means vector quantization algorithm and the GARCH parameters are estimated using only the signal's peak energy in each subband, as described in Section IV. For each of the sources, $K = 10$ linearly spaced subbands are considered and for the signal-absence state the covariance matrix is set to $\sigma^2_{\min,k} I$, where $\sigma^2_{\min,k}$ is 40 dB less than the higher averaged energy in the kth subband. Furthermore, in each subband, only frames in which the energy is within 40 dB of the peak energy (in the same subband) are considered for training the GMM.

The proposed algorithm is compared with the mmse estimator proposed in [4]. The latter algorithm assumes a single-band GMMs for both signals (i.e., with $K = 1$) and is referred to in the following as the GMM-based algorithm. This model is trained using the same training sets using the k-means algorithm.[4] For each of the algorithms, four, eight, and 16 states are considered for the GMM, while the GARCH model is trained with up to eight states per subband (excluding the signal absence state).

The performance evaluation in our study includes objective quality measures, a subjective study of waveforms and spectrograms, and informal listening tests. The first quality measure is the segmental signal-to-noise ratio (SNR) (in the time domain) which is defined in dB by [44] , as shown in (32) at the bottom of the page, where $\mathcal{H}_1$ represents the set of frames which contain the desired signal, $|\mathcal{H}_1|$ denotes the number of elements in $\mathcal{H}_1$, $N = 512$ is the number of samples per frame, and the operator $\mathcal{T}$ confines the SNR in each frame to a perceptually meaningful range between -10 and 35 dB. The second quality measure is log-spectral distortion (LSD) which is defined in dB by [45]

$$\text{LSD} = \frac{1}{L}\sum_{\ell=0}^{L-1}\left\{\frac{1}{N/2+1} \times \sum_{f=0}^{N/2}\left[10\log_{10}\mathcal{C}\mathbf{s}(\ell,f) - 10\log_{10}\mathcal{C}\hat{\mathbf{s}}(\ell,f)\right]^2\right\}^{\frac{1}{2}} \tag{33}$$

where $\mathbf{s}(\ell, f)$ denotes the fth element of the spectral vector $\mathbf{s}(\ell)$ (i.e., f denotes the frequency-bin index), $\mathcal{C}x \triangleq \max\{|x|^2, \epsilon\}$ is a spectral power clipped such that the log-spectrum dynamic range is confined to about 50 dB, that is, $\epsilon = 10^{-50/10} \times \max_{\ell,f}\{|\mathbf{s}(\ell,f)|^2\}$. Although the Segmental SNR and the LSD are common measures for speech enhancement, for the application of source separation, it was proposed in [46], [47] to measure the signal to interference ratio (SIR). For source s_1 we may write

$$\hat{s}_1 = \zeta_1 s_1 + \zeta_2 s_2 + d. \tag{34}$$

Accordingly, the Segmental SIR for $\hat{s}_1$ is defined in dB as follows:

$$\text{SegSIR} = \sum_{\ell}\mathcal{T}\left\{10\log_{10}\frac{\zeta_1(\ell)^2\sum_{n=0}^{N-1}s_1^2(n+\ell N/2)}{\zeta_2(\ell)^2\sum_{n=0}^{N-1}s_2^2(n+\ell N/2)}\right\} \tag{35}$$

where the parameters $\zeta_1(\ell)$ and $\zeta_2(\ell)$ are calculated for each segment as specified in [46].

The above-mentioned measures attempt to evaluate the averaged performance of the algorithm. The proposed classification and estimation approach enables one to control the tradeoff between the level of residual interference resulting from false detection of the desired signal and signal distortion resulting mainly from missed detection. To measure this tradeoff while applying the algorithm on a subband basis, we propose to measure the distortion of the estimated signal and the amount of interference reduction. Now let $\mathcal{H}_1$ and $\mathcal{H}_0$ denote the sets of (subband) frames which contain the desired signal and in which the desired signal is absent, respectively. The signal distortion, denoted as LSD_{H_1}, is evaluated using the LSD formulation (33) and averaged only over $\mathcal{H}_1$. The interference reduction is evaluated in dB by [48]

$$\text{IR}_{H_0} = 10\log_{10}\frac{\sum_{\ell\in\mathcal{H}_0}\|\hat{\mathbf{s}}_1(\ell)\|^2}{\sum_{\ell\in\mathcal{H}_0}\|\mathbf{s}_2(\ell)\|^2}. \tag{36}$$

B. Simulation Results

For evaluating the contribution of the proposed codebook (i.e., GARCH model for speech and GMM for music), the results obtained by using the proposed model are first compared with the results obtained by using the GMM-based algorithm. Since the GMM-based algorithm employs an mmse estimator, the proposed algorithm was applied in this experiment using constant cost parameters. As shown in Section II-A, this also yields mmse estimation. Fig. 3 shows quality measures as a function of the number of GARCH states.[5] These results are shown for 4-, 8- or 16-state GMM used for the music signal. For comparison, the results obtained by using the GMM-based

[4]Note that the k-means algorithm was used in our simulations to simplify the implementation, both in the proposed algorithm and in the reference algorithm. In practice, the EM algorithm may be used to train the GMM models as proposed in [4].

[5]The *improvements* in SegSNR and SegSIR are obtained by subtracting the initial values calculated for the mixed signals from those calculated for the processed signals.

$$\text{SegSNR} = \frac{1}{|\mathcal{H}_1|}\sum_{\ell\in\mathcal{H}_1}\mathcal{T}\left\{10\log_{10}\frac{\sum_{n=0}^{N-1}s^2(n+\ell N/2)}{\sum_{n=0}^{N-1}[s(n+\ell N/2)-\hat{s}(n+\ell N/2)]^2}\right\} \tag{32}$$

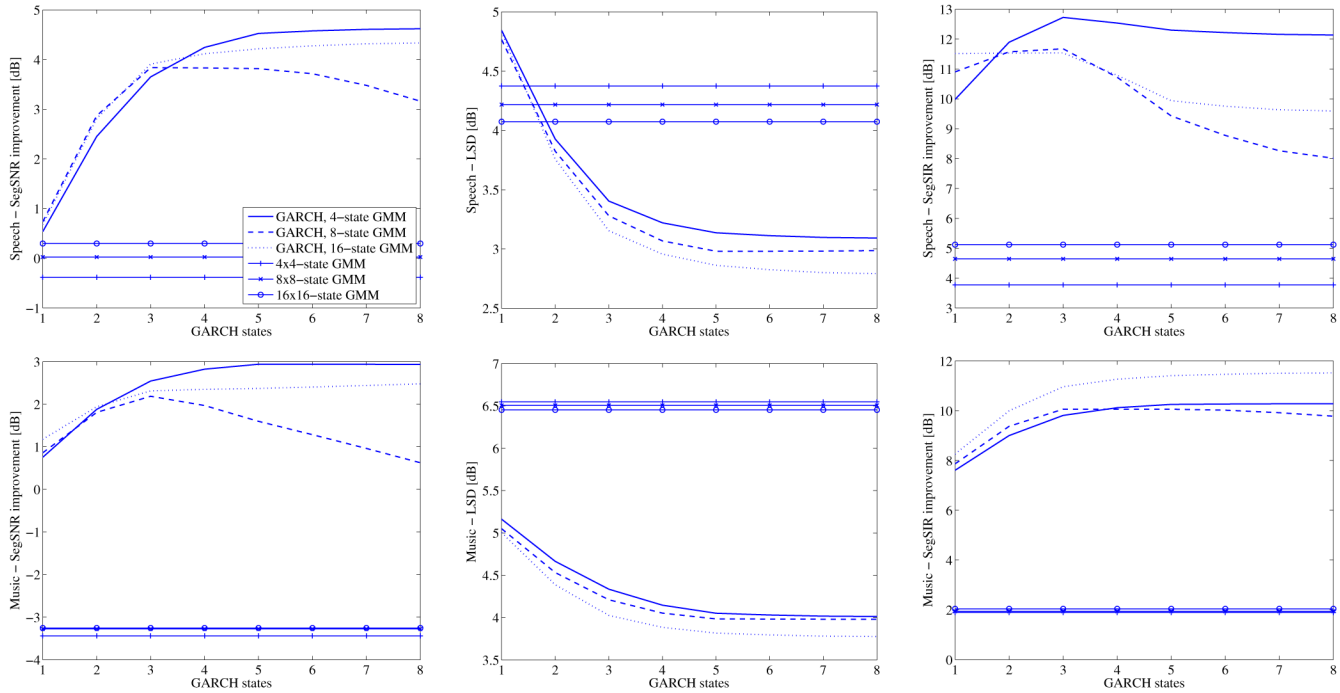

Fig. 3. Quality measures for mmse estimation as functions of the number of GARCH states. The results (with different numbers of GMM states for the music signal) are compared with the GMM-based algorithm. Upper row: results for speech signals; lower row: results for music signals. Columns (from left to right): SegSNR improvement, LSD, and SegSIR improvement.

algorithm are shown with 4, 8, and 16 states (for both signals). Note that for each algorithm, different number of subbands is considered, and different statistical model. However, the signal estimate in both cases is in the sense of mmse. It can be seen that employing a GARCH model for the speech signal significantly improves the separation results, and sometimes even using a single-state GARCH model outperforms the GMM modeling with up to 16 states. Moreover, it can be seen that excluding the SegSIR measure for speech signals, the performances are improved monotonically with the growth of the number of GARCH states (except for some cases with eight-state GMM). However, the significant improvement is obtained by using up to five states for the GARCH model with 4- or 16- state GMM for the music. Informal listening tests verify that increasing the number of GARCH states from one to 3 or 5, significantly improves the reconstructed signals and particularly the perceptual quality of the speech signal. Using three (or more) states for the speech model results in improved signals' quality compared to using the GMM for both the speech and the music signals. The GMM-based algorithm preserves mainly low frequencies of the music signal and the residual speech components sound somewhat scrappy. The proposed approach results in a more natural music signal which consists of higher range of frequencies. The residual speech signal also sounds more natural.

It is important to note that in general the proposed GARCH modeling requires higher computational complexity than the GMM modeling, since the covariance matrices need to be estimated recursively as described in Section III. However, the GARCH modeling requires smaller number of parameters and smaller number of states which reduce the computational complexity of the separation algorithm. While in the GMM model each state requires N parameters for the corresponding covariance matrix, in the GARCH case only three parameters are needed in each subband. If we consider similar separation technique, e.g., the mmse estimation (3), than using the GARCH modeling employs about four states for the speech signal while in the GMM case about 8 or 16 states are required to achieve reasonable separation. Therefore, in this case, the additional computations for the covariances estimation is compensated by less computations for the separation algorithm.

Next, we verify the performance of the proposed simultaneous classification and estimation method. As this method enables one to control the tradeoff between residual interference and signal distortion, we examine the influence of the cost parameters on these measures. The proposed algorithm was applied to the test set with different cost parameters. Fig. 4 shows the tradeoff between signal distortion and the reduction of the residual interference while examining the estimated speech signals. The averaged interfering reduction, IR_{H_0}, (in this case the reduction of the residual music) and the averaged speech distortion LSD_{H_1} are shown as functions of the false detection parameter for the speech signal $b_{1,\mathrm{f}}$ and for some values of the missed detection parameter. These results are evaluated using three-state GARCH model and eight-state GMM, and the simultaneous classification and estimation method. It is shown that when the false detection parameter increases, the level of residual interference decreases and the signal distortion increases. Therefore, for a specific application these parameters may be chosen to achieve a desired tradeoff between signal distortion and residual interference.

In Tables I and II, we provide quality measures for both types of signals using different sets of parameters. This test was

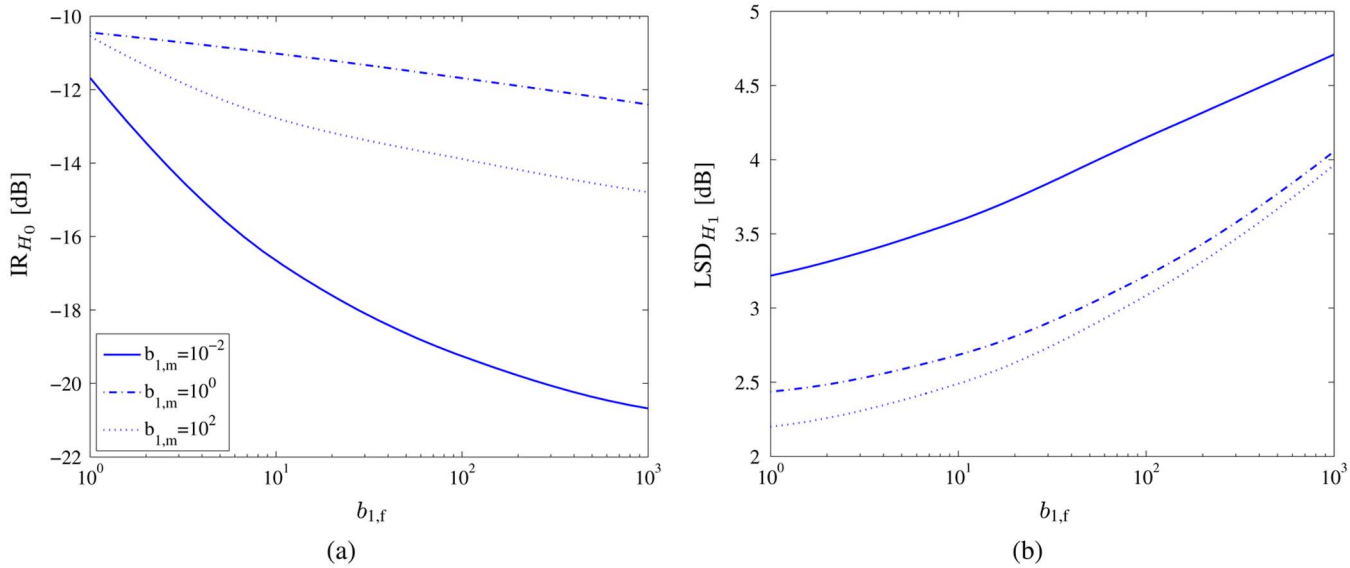

Fig. 4. Tradeoff between residual interference and signal distortion resulting from changing the false detection and missed detection parameters. (a) Residual music signal and (b) speech signal distortion.

TABLE I
AVERAGED QUALITY MEASURES FOR THE ESTIMATED SPEECH SIGNALS USING THREE-STATE GARCH MODEL AND EIGHT-STATE GMM

Parameters $[b_{1,m}, b_{1,f}, b_{2,m}, b_{2,f}]$	SegSNR improvement	SegSIR improvement	IR$_{H_0}$	LSD$_{H_1}$
$[1,1,1,1]$	3.67	11.67	-10.34	2.45
$[10^{-2},10^{2},10^{2},10^{-2}]$	**3.80**	13.83	**-15.39**	3.59
$[10^{2},10^{-2},10^{-2},10^{2}]$	3.54	11.00	-9.55	**2.04**
$[10^{-1},10^{1},10^{1},10^{-1}]$	3.76	**17.73**	-11.73	3.06
$[10^{2},10^{-2},10^{2},10^{-2}]$	3.68	11.44	-9.88	2.16

TABLE II
AVERAGED QUALITY MEASURES FOR THE ESTIMATED MUSIC SIGNALS USING THREE-STATE GARCH MODEL AND EIGHT-STATE GMM

Parameters $[b_{1,m}, b_{1,f}, b_{2,m}, b_{2,f}]$	SegSNR improvement	SegSIR improvement	IR$_{H_0}$	LSD$_{H_1}$
$[1,1,1,1]$	4.91	9.81	-7.27	3.04
$[10^{-2},10^{2},10^{2},10^{-2}]$	**5.46**	9.77	-5.95	3.35
$[10^{2},10^{-2},10^{-2},10^{2}]$	4.72	**10.05**	**-8.91**	**2.84**
$[10^{-1},10^{1},10^{1},10^{-1}]$	5.21	9.75	-6.27	3.25
$[10^{2},10^{-2},10^{2},10^{-2}]$	4.89	9.98	-7.47	2.91

conducted also for the whole test set using the simultaneous classification and estimation approach. It can be seen that by using different parameters, improved performance may be achieved compared to using equal parameters (i.e., using mmse estimation). However, as expected, different parameters would be needed to achieve the best performances in the sense of different quality measures. Specifically, in case of speech signals, the higher interference-reduction is achieved with the parameters (from the tested sets of parameters) which corresponds to the highest distortion. On the other hand, the lowest distortion is obtained with the lowest amount of interference reduction.

Figs. 5 and 6 demonstrate the separation of a specific mixture of speech and piano signals. The speech waveform, the piano waveform, and their mixture are shown in Fig. 5, and Fig. 6 shows the separated signals resulting from an eight-state GMM-based algorithm and from the proposed simultaneous classification and separation approach (using three-state GARCH model

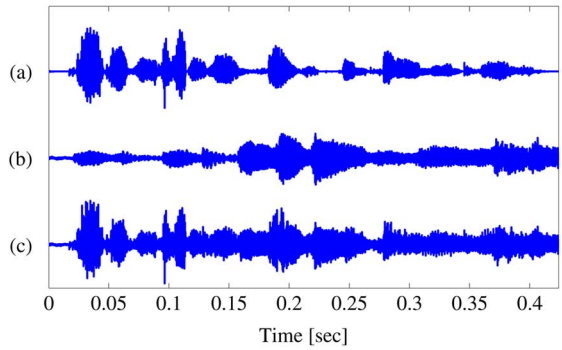

Fig. 5. Original and mixed signals. (a) Speech signal: "Draw every outer line first, then fill in the interior": (b) piano signal (*Für Elise*), (c) mixed signal.

for the speech signal, eight-state GMM for the piano signal, $b_{1,\mathrm{m}} = b_{2,\mathrm{f}} = 5$, and $b_{1,\mathrm{f}} = b_{2,\mathrm{m}} = 15$). It can be seen that for this particular mixture, by estimating the speech signal the proposed algorithm results in higher attenuation of the piano signal, and the estimation of the piano signal preserves more energy of the desired signal, especially at its second half.

VI. CONCLUSION

We have proposed a new approach for single-channel audio source separation of acoustic signals, which is based on classifying the mixed signal into codebooks, and estimating the subsources. Unlike other classical methods which apply estimation alone, or distinctive operations of classification and estimation, in our method both operations are designed simultaneously, or the estimator is designed to allow a compensation for erroneous classification. In addition, a new codebook is proposed for speech signals in the STFT domain based on the GARCH model. Accordingly, less restrictive pdf's are enabled in the STFT domain compared to GMM or AR-based model. Experimental results on mixtures of speech and piano music signals have yielded an improved source separation performance

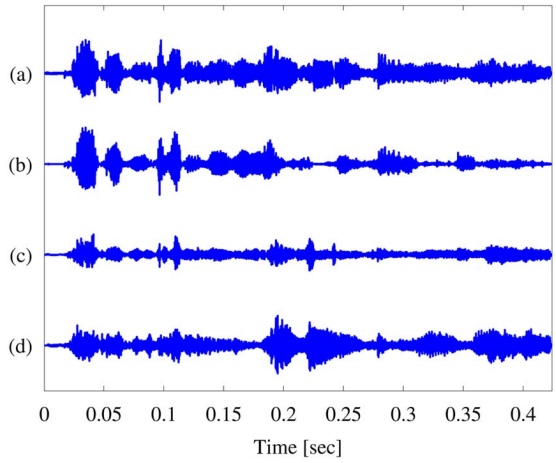

Fig. 6. Separation of speech and music signals. (a) speech signal reconstructed by using the GMM-based algorithm (SegSNR improvement $= 0.76$LSD $= 3.77$, SegSIR improvement $= 1.29$). (b) Speech signal reconstructed using the proposed approach (SegSNR improvement $= 2.46$LSD $= 3.56$, SegSIR improvement $= 8.61$). (c) Piano signal reconstructed by using the GMM algorithm (SegSNR improvement $= -2.77$, LSD $= 4.34$, SegSIR improvement $= 2.50$). (d) Piano signal reconstructed using the proposed approach (SegSNR improvement $= 0.32$, LSD $= 3.19$, SegSIR improvement $= 4.79$).

compared to using GMM for both signals, even when using a smaller number of states. In addition, applying a simultaneous classification and estimation approach enables one to control the tradeoff between signal distortion and residual interference.

The proposed classification and estimation method may be advantageously utilized with other codebooks for different types of signals. However, the selection of the optimal parameters in the general case may be codebook- as well as application-dependent and may be a subject for further research. Furthermore, the GARCH modeling for speech signals may be combined with various statistical models for the music signals, other than GMM, such as mixture of AR or HMM with AR subsources.

Appendix A
Derivation of (10)

By setting the derivative of $\sum_{\bar{i},\bar{j}} p(\bar{i},\bar{j}) r_{ij}^{\bar{i}\bar{j}}(\mathbf{x},\hat{\mathbf{s}}_1)$ in (9) to zero we obtain

$$0 = \sum_{\bar{i}\bar{j}} b_{ij}^{\bar{i}\bar{j}} p(\bar{i}\bar{j}) \left[\hat{\mathbf{s}}_{1,ij} \int p(\mathbf{x}\,|\,\mathbf{s}_1,\bar{j}) p(\mathbf{s}_1\,|\,\bar{i}) d\mathbf{s}_1 - \int \mathbf{s}_1\, p(\mathbf{x}\,|\,\mathbf{s}_1,\bar{j}) p(\mathbf{s}_1\,|\,\bar{i}) d\mathbf{s}_1 \right] \quad (37)$$

where

$$p(\mathbf{x}\,|\,\mathbf{s}_1,\bar{j}) p(\mathbf{s}_1\,|\,\bar{i}) = p(\mathbf{x}\,|\,\mathbf{s}_1,\bar{i},\bar{j}) p(\mathbf{s}_1\,|\,\bar{i},\bar{j}) = p(\mathbf{x}\,|\,\bar{i},\bar{j}) p(\mathbf{s}_1\,|\,\mathbf{x},\bar{i},\bar{j}). \quad (38)$$

Substituting (38) into (37) we obtain

$$0 = \sum_{\bar{i}\bar{j}} b_{ij}^{\bar{i}\bar{j}} p(\bar{i}\bar{j}) [\hat{\mathbf{s}}_{1,ij} p(\mathbf{x}\,|\,\bar{i},\bar{j}) - p(\mathbf{x}\,|\,\bar{i},\bar{j}) E\{\mathbf{s}_1\,|\,\mathbf{x},\bar{i},\bar{j}\}] \quad (39)$$

and accordingly

$$\hat{\mathbf{s}}_{1,ij} = \frac{\sum_{\bar{i}\bar{j}} b_{ij}^{\bar{i}\bar{j}} p(\mathbf{x}\,|\,\bar{i},\bar{j}) p(\bar{i},\bar{j}) W_{\bar{i}\bar{j}}\,\mathbf{x}}{\sum_{\bar{i}\bar{j}} b_{ij}^{\bar{i}\bar{j}} p(\mathbf{x}\,|\,\bar{i},\bar{j}) p(\bar{i},\bar{j})}. \quad (40)$$

Appendix B
Derivation of (11)

The average risk is given by

$$r_{ij}^{\bar{i}\bar{j}}(\mathbf{x},\hat{\mathbf{s}}_1) = \int C_{ij}^{\bar{i}\bar{j}}(\mathbf{s}_1,\hat{\mathbf{s}}_1) p(\mathbf{x}\,|\,\mathbf{s}_1,\bar{j}) p(\mathbf{s}_1\,|\,\bar{i}) d\mathbf{s}_1 = \int b_{ij}^{\bar{i}\bar{j}} \|\mathbf{s}_1 - \hat{\mathbf{s}}_1\|_2^2\, p(\mathbf{x},\mathbf{s}_1\,|\,\bar{i},\bar{j}) d\mathbf{s}_1. \quad (41)$$

To simplify the notation, we assume in this Appendix that the active states of both signals are known, so we may omit the indices $\{\bar{i},\bar{j}\}$. Furthermore, we use $\mathbf{s}$ to denote $\mathbf{s}_1$ and we assume diagonal covariance matrices $\Sigma_1 = \mathrm{diag}\{\sigma_1^2(1),\sigma_1^2(2),\ldots,\sigma_1^2(N)\}$ and $\Sigma_2 = \mathrm{diag}\{\sigma_2^2(1),\sigma_2^2(2),\ldots,\sigma_2^2(N)\}$. Following these notations we obtain

$$\int \|\mathbf{s}\|_2^2 p(\mathbf{x},\mathbf{s}) d\mathbf{s} = \sum_f \left\{ \int |\mathbf{s}(f)|^2 p(\mathbf{x}(f),\mathbf{s}(f)) d\mathbf{s}(f) \times \prod_{f' \neq f} \int p(\mathbf{x}(f'),\mathbf{s}(f')) d\mathbf{s}(f') \right\} \quad (42)$$

where in this Appendix $\mathbf{s}(f)$ and $\mathbf{x}(f)$ denote the fth elements of vectors $\mathbf{s}$ and $\mathbf{x}$, respectively (i.e., f denotes the frequency-bin index). Let $\lambda(f) \triangleq (\sigma_1^2(f)^{-1} + \sigma_2^2(f)^{-1})^{-1}$, let $\xi(f) \triangleq \sigma_1^2(f)/\sigma_2^2(f)$, let $\gamma(f) \triangleq |\mathbf{x}(f)|^2/\sigma_2^2(f)$, and let $\upsilon(f) \triangleq \xi(f)\gamma(f)/(1+\xi(f))$. By integrating over both the real and imaginary parts of $\mathbf{s}(f)$ and using ([49], (3.462.2)) we obtain

$$\int |\mathbf{s}(f)|^2 p(\mathbf{x}(f),\mathbf{s}(f)) d\mathbf{s}(f) = \frac{\xi(f)(1+\upsilon(f))}{\pi(1+\xi(f))^2} \times \exp\left\{ -\frac{\gamma(f)}{1+\xi(f)} \right\} \quad (43)$$

and

$$\int p(\mathbf{x}(f'),\mathbf{s}(f')) d\mathbf{s}(f') = p(\mathbf{x}(f')) = \frac{\exp\left\{ -\frac{\gamma(f')}{1+\xi(f')} \right\}}{\pi\sigma_2^2(f')(1+\xi(f'))}. \quad (44)$$

Let $\Xi \triangleq \mathrm{diag}\{\xi(1),\xi(2),\ldots,\xi(N)\}$ and $V \triangleq \mathrm{diag}\{\upsilon(1),\upsilon(2),\ldots,\upsilon(N)\}$. Substituting (43) and (44) into (42) we obtain

$$\int \|\mathbf{s}\|_2^2\, p(\mathbf{x},\mathbf{s})\, d\mathbf{s} = \sum_f \left\{ \frac{\xi(f)\,(1+\upsilon(f))}{\pi\,(1+\xi(f))^2} \exp\left(\frac{-\gamma(f)}{1+\xi(f)} \right) \right.$$

$$\times \prod_{f' \neq f} \frac{1}{\pi\, \sigma_2^2(f')\,(1+\xi(f'))} \exp\left(\frac{-\gamma(f')}{1+\xi(f')}\right)\Bigg\}$$
$$= \frac{\mathbf{1}^T \Sigma_1 (I+V)(I+\Xi)^{-1}\mathbf{1}}{\pi^N |\Sigma_2 (I+\Xi)|} \times \exp\{-\mathbf{x}^H(\Sigma_1+\Sigma_2)^{-1}\mathbf{x}\}. \quad (45)$$

Let subscripts $_R$ and $_I$ denote the real and imaginary parts of a complex-valued variable, respectively, and let $g_{ij}(f)$ denote the fth diagonal element of matrix G_{ij}. Then, using ([49], (3.462.2)) we obtain

$$\int (\hat{\mathbf{s}}^H \mathbf{s} + \mathbf{s}^H \hat{\mathbf{s}}) p(\mathbf{x},\mathbf{s}) d\mathbf{s}$$
$$= 2\int \left(\hat{\mathbf{s}}_R^T \mathbf{s}_R + \mathbf{s}_I^T \hat{\mathbf{s}}_I\right) p(\mathbf{x},\mathbf{s}) d\mathbf{s}$$
$$= 2\sum_f \Bigg\{ \int [\mathbf{x}_R(f)\mathbf{s}_R(f) + \mathbf{x}_I(f)\mathbf{s}_I(f)] p(\mathbf{x}(f),\mathbf{s}(f)) d\mathbf{s}(f) \times g_{ij}(f) \prod_{f' \neq f} \int p(\mathbf{x}(f'),\mathbf{s}(f')) d\mathbf{s}(f') \Bigg\}$$
$$= 2\sum_f \frac{g_{ij}(f) v(f) \sigma_2^2(f)}{\pi} \prod_{f' \neq f} \frac{\exp\left\{-\frac{\gamma(f')}{1+\xi(f')}\right\}}{\pi \sigma_2^2(f')(1+\xi(f'))}$$
$$= 2\frac{\mathbf{1}^T G_{ij} \Sigma_2 V \mathbf{1}}{\pi^N |\Sigma_2 (I+\Xi)|} \exp\left\{-\mathbf{x}^H(\Sigma_1+\Sigma_2)^{-1}\mathbf{x}\right\}. \quad (46)$$

Finally,

$$\int p(\mathbf{x},\mathbf{s}) d\mathbf{s} = p(\mathbf{x}) = \frac{\exp\{-\mathbf{x}^H(\Sigma_1+\Sigma_2)^{-1}\mathbf{x}\}}{\pi^N |\Sigma_1+\Sigma_2|}. \quad (47)$$

Substituting (45)–(47) into (42) and using $W = \Xi(I+\Xi)^{-1}$, we obtain

$$r_{ij}(\mathbf{x},\hat{\mathbf{s}}_1) = b_{ij}^{\vec{ij}}\, p(\mathbf{x})[\mathbf{x}^H(W^2 - 2WG_{ij})\mathbf{x} + \mathbf{1}^T \Sigma_2 W \mathbf{1}]$$
$$= \frac{b_{ij}}{\pi^N |\Sigma_1+\Sigma_2|} \exp\{-\mathbf{x}^H(\Sigma_1+\Sigma_2)^{-1}\mathbf{x}\} \times [\mathbf{x}^H(W^2 - 2WG_{ij})\mathbf{x} + \mathbf{1}^T \Sigma_2 W \mathbf{1}]. \quad (48)$$

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their valuable comments and helpful suggestions.

REFERENCES

[1] G.-J. Jang, T.-W. Lee, and Y.-H. Oh, "Single-channel signal separation using time-domain basis functions," *IEEE Signal Process. Lett.*, vol. 10, no. 6, pp. 168–171, Jun 2003.

[2] M. V. S. Shashanka, B. Raj, and P. Smaragdis, "Sparse overcomplete decomposition for single channel speaker separation," in *Proc. 32nd IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP'07*, Honolulu, HI, Apr. 2007, pp. 641–644.

[3] L. Benaroya and F. Bimbot, "Wiener based source seperation with HMM/GMM using a single sensor," in *Proc. 4th Int. Symp. Independent Compon. Anal. Blind Signal Separation (ICA)*, Nara, Japan, Apr. 2003, pp. 957–961.

[4] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 14, no. 1, pp. 191–199, Jan. 2006.

[5] L. Benaroya, F. Bimbot, G. Gravier, and R. Gribonval, "Experiments in audio source separation with one sensor for robust speech recognition," *Speech Commun.*, vol. 48, no. 7, pp. 848–854, Jul. 2006.

[6] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, New Paltz, NY, Oct. 2005, pp. 90–93.

[7] L. Benaroya, R. Blouet, C. Févotte, and I. Cohen, "Single sensor source separation based on wiener filtering and multiple window STFT," in *Proc. Int. Workshop Acoust. Echo Noise Control, IWAENC'06*, Paris, France, Sep. 2006, pp. 1–4, paper 52,.

[8] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1564–1578, Jul. 2007.

[9] Y. Ephraim, D. Malah, and B.-H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 12, pp. 1846–1856, Dec. 1989.

[10] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden marks models," *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 725–735, Apr. 1992.

[11] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Networks*, vol. 5, no. 14, pp. 1135–1150, Sep. 2004.

[12] S. Srinivasan, J. Smuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement," in *Proc. 30nd IEEE Int. Conf. Acoust., Speech Signal Process., ICASSP'05*, Philadelphia, PA, Mar. 2005, vol. 1, pp. 1077–1080.

[13] S. Srinivasan, J. Smuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 163–173, Jan. 2006.

[14] J.-F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.

[15] M. J. Reyes-Gomez, D. P. W. Ellis, and N. Jojic, "Multiband audio modelong for single-channel acoustic source separation," in *Proc. 29st IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP'04*, Montreal, QC, Canada, May 2004, pp. 641–644.

[16] I. Cohen, "Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models," *Signal Process.*, vol. 86, no. 4, pp. 698–709, Apr. 2006.

[17] I. Cohen, "Modeling speech signals in time-frequency domain using GARCH," *Signal Process.*, vol. 84, no. 12, pp. 2453–2459, Dec. 2004.

[18] A. Abramson and I. Cohen, "Recursive supervised estimation of a Markov-switching GARCH process in the short-time fourier transform domain," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3227–3238, Jul. 2007.

[19] A. Abramson and I. Cohen, "Markov-switching GARCH model and application to speech enhancement in subbands," in *Proc. Int. Workshop Acoust. Echo Noise Control, IWAENC'06*, Paris, France, Sep. 2006, pp. 1–4, paper 7,.

[20] M. Abdolahi and H. Amindavar, "GARCH coefficients as feature for speech recognition in persian isolated digit," in *Proc. 30th IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP'05*, Philadelphia, PA, May 2005, pp. I.957–I.960.

[21] R. Tahmasbi and S. Rezaei, "A soft voice activity detection using GARCH filter and variance gamma distribution," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 5, no. 4, pp. 1129–1134, May 2007.

[22] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *J. Econometrics*, vol. 31, no. 3, pp. 307–327, 1986.

[23] J. D. Hamilton, *Time Series Analysis*. Princeton, NJ: Princeton Univ. Press, 1994.

[24] I. Cohen, , J. Benesty, S. Makino, and J. Chen, Eds., "From volatility modeling of financial time-series to stochastic modeling and enhancement of speech signals," in *Speech Enhancement*. New York: Springer, 2005, ch. 5, pp. 97–114.

[25] F. D. Neeser and J. L. Massey, "Proper complex random process with applications to information theory," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1293–1302, Jul. 1993.

[26] D. G. Manokis, V. K. Ingle, and S. M. Kogon, *Statistical and Adaptive Signal Process.: Spectral Estimation, Signal Modeling, Adaptive Filtering, and Array Processing*. Boston, MA: McGraw-Hill, 2000.

[27] D. Middleton and F. Esposito, "Simultaneous optimum detection and estimation of signals in noise," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 3, pp. 434–444, May 1968.
[28] A. Fredriksen, D. Middleton, and D. Vandelinde, "Simultaneous signal detection and estimation under multiple hypotheses," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 5, pp. 607–614, 1972.
[29] A. Abramson and I. Cohen, "Simultaneous detection and estimation approach for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2348–2359, Nov. 2007.
[30] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.
[31] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 334–341, Jul. 2003.
[32] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Nashua, NH: Athena Scientific, 1999.
[33] A. P. Dempster, N. M. Larid, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.
[34] G. A. F. Seber, *Multivariate Observations*. New York: Wiley, 1984.
[35] Y. Ephraim and D. Malah, "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.
[36] A. Abramson and I. Cohen, "On the stationarity of GARCH processes with Markov switching regimes," *Econometric Theory*, vol. 23, no. 3, pp. 485–500, 2007.
[37] F. Klaassen, "Improving GARCH volatility forecasts with regimeswitching GARCH," *Empirical Economics*, vol. 27, no. 2, pp. 363–394, Mar. 2002.
[38] M. Haas, S. Mittnik, and M. S. Paolella, "A new approach to Markov switching GARCH models," *J. Financial Econometrics*, vol. 2, no. 4, pp. 493–530, Autumn, 2004.
[39] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.
[40] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary environments," *Signal Process.*, vol. 81, pp. 2403–2418, Nov. 2001.
[41] H. Bourlard and S. Dupont, "Subband-based speech recognition," in *Proc. 22nd IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP'97*, Munich, Germany, Apr. 1997, pp. 1251–1254.
[42] S. Okawa, E. Bocchieri, and A. Potamianos, "Multi-band speech recognition in noisy environments," in *Proc. 23rd IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP-98*, Seattle, WA, USA, May 1998, pp. 641–644.
[43] J. S. Garofolo, Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database National Inst. of Stand. and Technol. (NIST), Gaithersburg, MD, Tech. Rep., Dec. 1988.
[44] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Meaasures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
[45] I. Cohen and S. Gannot, , J. Benesty, M. M. Sondhi, and Y. Huang, Eds. , "Spectral enhancement methods," in *Springer Handbook of Speech Processing*. New York: Springer, 2007, ch. 45.
[46] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.
[47] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte, "Proposals for performance measurement in source separation," in *Proc. 4th Int. Symp. Independent Compon. Anal. Blind Signal Separation (ICA)*, Nara, Japan, Apr. 2003, pp. 763–768.
[48] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.
[49] I. S. Gradshteyn and I. M. Ryzhik, , A. Jefferey and D. Zwillinger, Eds., *Table of Integrals, Series, and Products*, 6th ed. New York: Academic, 2000.

Ari Abramson (S'06) received the B.Sc. degree in electrical engineering from the Tel-Aviv University, Tel-Aviv, Israel, in 2002 and the M.Sc. and Ph.D. degrees in electrical engineering from the Technion—Israel Institute of Technology, Haifa, Israel, in 2007 and 2008, respectively.

From 1993 to 2004, he served as a combat copilot in the Israeli Air Force, and since 2004 he has been a Flight-Test Engineer in reserve duty. His research interests are statistical signal processing, speech enhancement and detection, and estimation theory.

Dr. Abramson received the Wolf foundation excellence award in 2005 and the Best Student Paper Award at the International Workshop on Acoustic, Echo and Noise Control in 2006.

Israel Cohen (M'01–SM'03) received the B.Sc. (summa cum laude), M.Sc., and Ph.D. degrees in electrical engineering from the Technion—Israel Institute of Technology, Haifa, Israel, in 1990, 1993, and 1998, respectively.

From 1990 to 1998, he was a Research Scientist with RAFAEL Research Laboratories, Israel Ministry of Defense, Haifa. From 1998 to 2001, he was a Postdoctoral Research Associate with the Computer Science Department, Yale University, New Haven, CT. In 2001, he joined the Electrical Engineering Department of the Technion, where he is currently an Associate Professor. His research interests are statistical signal processing, analysis and modeling of acoustic signals, speech enhancement, noise estimation, microphone arrays, source localization, blind source separation, system identification, and adaptive filtering.

Dr. Cohen received in 2005 and 2006 the Technion Excellent Lecturer awards. He served as Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and IEEE SIGNAL PROCESSING LETTERS, and as Guest Editor of a special issue of the *EURASIP Journal on Advances in Signal Process.* on Advances in Multimicrophone Speech Processing and a special issue of the *EURASIP Speech Communication Journal* on Speech Enhancement. He is a coeditor of the Multichannel Speech Processing section of the *Springer Handbook of Speech Processing* (Springer, 2007).