Markov-Switching GARCH Models and Applications to Digital Speech Processing

Ari Abramson

Markov-Switching GARCH Models and Applications to Digital Speech Processing

Research Thesis

As Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy

Ari Abramson

Submitted to the Senate of the Technion—Israel Institute of Technology

Tevet 5768 Haifa December 2007

The Research Thesis was Done Under the Supervision of Associate Professor Israel Cohen in the Department of Electrical Engineering.

Acknowledgement

I wish to express my deep gratitude and appreciation to Prof. Israel Cohen for his guidance and dedicated supervision. Thank for your professional support, for your encouragement to perfection, and for many valuable suggestions throughout all the stages of this research.

I would also like to thank Kuti Avargel for many fruitful discussions and Dr. Emanuël Habets for valuable discussions and for giving me the opportunity to expand my research to speech dereverberation.

Special thanks to my parents, Miri and Moshe and to my beloved Efrat who encouraged and supported me through the whole way.

The Generous Financial Help of The Technion, The Israel Science Foundation (Grant no. 1085/05), and The European Commission's IST Program Under Project Memories is Gratefully Acknowledged.

Contents

List of Figures

List of Tables

Abstract

This dissertation addresses theory and applications of generalized autoregressive conditional heteroscedasticity (GARCH) models with Markov regimes for digital speech processing. The GARCH model is widely-used in the field of econometrics for volatility forecast derivation of econometric rates, and it was recently proposed in the field of signal processing for applications such as speech enhancement, speech recognition, and voice activity detection. GARCH models explicitly parameterize the time-varying volatility by using both past conditional variances and past squared innovations (prediction errors), while taking into account excess kurtosis (i.e., heavy tailed distribution) and volatility clustering.

In this thesis, we develop a new statistical model for nonstationary signals in the joint time-frequency domain based on GARCH formulation with Markov regimes. The proposed model exploits the advantages of both the conditional heteroscedasticity structure of GARCH models and the time-varying characteristics of hidden Markov chains. The main motivation for this research is spectral modeling of speech signals for hands-free communication applications such as speech enhancement, nonstationary noise reduction, dereverberation, and audio source separation.

We analyze the asymptotic stationarity of Markov-switching GARCH (MS-GARCH) processes in the general case of (p, q)-order GARCH models with finite-state Markov chains. Necessary and sufficient conditions for asymptotic wide-sense stationarity are developed for several model formulations which are known in the literature. The properties of the proposed model are investigated and algorithms are developed for conditional variance, as well as for signal estimation in noisy environments. The proposed model with the corresponding estimation algorithms are shown to be useful for applications of speech enhancement and speech dereverberation. In addition, a state smoothing algorithm is

developed for the sequence of active states estimation. Furthermore, a new formulation for the speech enhancement problem is proposed in this thesis, which incorporates simultaneous operations of detection and estimation. A detector for speech presence in the short-time Fourier transform domain is combined with an estimator, which jointly minimizes a cost function that takes into account both detection and estimation errors. We show that the proposed simultaneous detection and estimation approach enables greater noise reduction than estimation only approach, without further degrading the speech signal.

A simultaneous classification and estimation approach together with GARCH modeling is employed for developing an algorithm for single-sensor audio source separation. We show that for mixtures of speech and music signals, an improved source separation can be achieved compared to using Gaussian mixture model for both signals. Moreover, cost parameters enable one to control the trade-off between missed and false detection of the desired signal, and correspondingly the trade-off between signal distortion and residual interference.

Notation

x	scalar / time-domain signal
$\boldsymbol{x}$	column vector
X_{tk} , $X(t,k)$	time-frequency coefficient
$\{A\}_{ij}$, a_{ij}	the (i,j) element of matrix A
A^{-1}	matrix inverse
diag$\{\boldsymbol{x}\}$	diagonal matrix with the vector$\boldsymbol{x}$ on its diagonal
eig$\{A\}$	the spectrum of matrix A
tr$\{\cdot\}$	trace
$\lvert x\rvert$	absolute value
$\lvert A\rvert$	determinant
$(\cdot)^T$, $(\cdot)'$	transpose operation
$(\cdot)^H$	Hermitian
$(\cdot)^*$	complex conjugate
$p(\cdot)$	probability / probability density
$E\{\cdot\}$	expectation
$Var\{\cdot\}$	variance
$Cov\{\cdot\}$	covariance
det$(\cdot)$	determinant
$I_\nu(\cdot)$	modified Bessel function of order ν
$J_\nu(\cdot)$	Bessel function of order ν
$\Gamma(\cdot)$	Gamma function
${}_1F_1\,(a;b;x)$	confluent hypergeometric function
$\rho(\cdot)$	spectral radius

$\mathbf{1}$	column vector of ones
0_m	$m \times m$ matrix of zeros
I_m	$m \times m$ identity matrix
$\otimes$	Kronecker product
$\odot$	term-by-term vector multiplication
$(\div)$	term-by-term vector division
∇	gradient
$\mathbb{R}^N$	N-demential real-valued vectors
$\mathbb{R}_+^N$	N-demential positive real-valued vectors
$\mathbb{C}^N$	N-demential complex-valued vectors
$\|\cdot\|$	Euclidian norm
$(\cdot)_R$	real part
$(\cdot)_I$	imaginary part

Abbreviations

AIR	Acoustic impulse response
AR	Autoregressive
ARCH	Autoregressive conditional heteroscedasticity
ARMA	Autoregressive moving average
DPC	Direct path compensation
DSB	Delay and sum beamformer
GARCH	Generalized autoregressive conditional heteroscedasticity
GMM	Gaussian mixture model
HMM	Hidden Markov model
HMP	Hidden Markov process
IMCRA	Improved minima-controlled recursive averaging
$\mathrm{IR}_{\mathcal{H}_0}$	Interference reduction
LRSVE	Late reverberant spectral variance estimator
LSA	Log spectral amplitude
LSD	Log spectral distortion
MAP	Maximum a posteriori
ML	Maximum likelihood
MMSE	Minimum mean-square error
MSE	Mean-square error
MS-GARCH / MSG	Markov-switching GARCH
MSTF-GARCH	Markov-switching time-frequency GARCH
NE	Noise estimator
OM-LSA	Optimally modified log-spectral amplitude

PESQ	Perceptual evaluation of speech quality
PSD	Power spectral density
QSA	Quadratic spectral amplitude
RIR	Room impulse response
SegSNR	Segmental signal-to-noise ratio
SegSIR	Segmental signal-to-interference ratio
SNR	Signal-to-noise ratio
STFT	Short-time Fourier transform
STSA	Short-term spectral amplitude
TF-GARCH	Time-frequency generalized autoregressive conditional heteroscedasticity
VAD	Voice activity detector

Chapter 1

Introduction

This dissertation addresses the theory of generalized autoregressive conditional heteroscedasticity (GARCH) models with Markov regimes and their applications to signal processing, and in particular to digital speech processing.

GARCH model explicitly parameterizes a time-varying volatility by using both recent conditional variances and recent squared innovations. This model is widely-used in the field of econometrics for the analysis and volatility forecasts in financial markets. Recently, this model was proposed for speech processing applications, such as speech enhancement, speech recognition, and voice activity detection. In this thesis we formulate a new complex-valued Markov-switching GARCH (MS-GARCH) model for nonstationary signals in the joint time-frequency domain. The MS-GARCH model exploits the advantages of both the conditional heteroscedasticity structure of GARCH models and the time-varying characteristics of hidden Markov chains. The basic motivation for our research is based on spectral modeling of speech signals, where examples for applications may include, e.g., noise reduction in communication systems (both background and transient noise), speech enhancement and dereverberation in hands-free communication, and audio source separation.

The thesis starts with an asymptotic stationarity analysis of MS-GARCH processes. In case of processes with time-varying variances, conditions for asymptotic wide-sense stationarity are useful to ensure a stable process, with a finite second-order moment. We then formulate a new complex-valued MS-GARCH model for nonstationary signals in the short-time Fourier transform (STFT) domain. The properties of the model are

investigated and algorithms are developed for causal, as well as noncausal estimation in noisy environment. The proposed model is shown to be useful for spectral modeling of speech signals for the applications of speech enhancement and speech dereverberation. A basic property of speech signals is that their expansion coefficients are sparse in the STFT domain. Therefore, a reliable detector may significantly improve performance in noisy environments. We propose a new formulation for the speech enhancement problem, which incorporates simultaneous operations of detection and estimation. This approach is applied to develop speech enhancement algorithms under stationary, as well as transient noise. A simultaneous classification and estimation approach together with GARCH modeling is employed to develop an algorithm for single-sensor audio source separation.

In this chapter we briefly describe scientific background for the main topics of this research and specify the structure of the thesis.

1.1 Markov-switching GARCH models

The GARCH model is widely-used in the field of econometrics, both by practitioners and by researchers [1–5]. The model represents a powerful tool for analysis and forecasting of volatility in financial markets. This model, first introduced by Bollerslev [1] as a generalization of the ARCH model [2], explicitly parameterizes the time-varying volatility by using both recent conditional variances and recent squared innovations. GARCH models preserve the persistence of the process volatility in the sense that small variations tend to follow small variations and large variations tend to follow large variations. Incorporating GARCH models with hidden Markov chains, where each state (regime) of the chain implies a different GARCH behavior, extends the dynamic formulation of the model and enables a better fit for a process with a more complex time-varying volatility structure [6–11]. However, a major drawback of such models is that estimating the volatility with switching-regimes requires knowledge of the entire history of the process, including the regime path. Consequently, Markov-switching ARCH models were proposed [12, 13], which avoid problems of path dependency in a noiseless environment. The conditional variance in ARCH models depends on previous observations only, so the Markov chain does not have to be known for constructing the conditional variance for a given regime.

In [6], a variant of MS-GARCH (MS-GARCH) model was introduced relying on the assumption that the conditional variance given current regime is dependent on the *expectation* of the previous conditional variances rather than their values. Accordingly, the conditional variance depends on some finite, state dependent, expected conditional variances via their conditional state probabilities. Klaassen [7] proposed modifying this model by manipulating the current regime and all available observations while evaluating the expectation of previous conditional variances. A different method for reducing the dependency of the conditional variance on past regimes has recently been proposed in [8]. Accordingly, a Markov chain governs the ARCH parameters while the autoregressive behavior of the conditional variance is subject to the assumption that past conditional variances are in the same regime as that of the current conditional variance. These variants of MS-GARCH models were developed for improved volatility forecasts of financial time-series under possible existence of shocks.

MS-GARCH processes, as well as the standard (single-state) GARCH process, are non-stationary as their second-order moments change recursively over time. However, if these processes are asymptotically wide-sense stationary then their variances are guarantied to be finite. A necessary and sufficient condition for the stationarity of a (single-regime) GARCH(p, q) process has been developed in [1]. A deep analysis of the probabilistic structure of MS-GARCH model is derived in [14] with conditions for the existence of moments of any order. In [15–17], stationarity analysis has been derived for some mixing models of conditional heteroscedasticity, and conditions for the asymptotic stationarity of some AR and ARMA models with Markov-regimes has been derived in [18–22]. However, for the MS-GARCH models, stationarity conditions are known in the literature only for some special cases. In [7], necessary (but not necessarily sufficient) conditions for stationarity are developed for the special case of two regimes and GARCH modeling of order $(1, 1)$. A necessary and sufficient stationarity condition has been developed in [8] for a specific MS-GARCH model formulation, but only in case of GARCH$(1, 1)$ behavior in each regime. We introduce a comprehensive approach for stationarity analysis of MS-GARCH models, which manipulates a backward recursion of the model's second-order moment. A recursive formulation of the state-dependent conditional variances is developed and the corresponding conditions for stationarity are obtained. In particular, we derive necessary

and sufficient conditions for the asymptotic wide-sense stationarity of two different variants of MS-GARCH processes, and obtain expressions for their asymptotic variances in the general case of m-state Markov chains and (p, q)-order GARCH processes.

Recently, GARCH models have been employed for modeling speech signals in the time-frequency domain [23–25], for speech recognition application [26], and for voice activity detection [27]. Speech signals in the STFT domain demonstrate both variability clustering and heavy tail behavior similarly to financial time-series [25]. Motivated by these characteristics, it was proposed to model the conditional variance of speech signals in the STFT domain by a complex GARCH model. This model has been shown useful for speech enhancement applications [23–25], but it relies on the assumption that the model parameters are time-invariant. It is commonly assumed in econometrics that the analyzed process is observed in a noiseless environment so that its past observations provide a complete specification of its current conditional variance, for any given regime. In our proposed MS-GARCH approach for speech modeling, the desired signal is generally observed in a noisy environment. Accordingly, we developed algorithm for conditional variance estimation which is based on iterating propagation and update steps with regime conditional probabilities. In addition, we developed state smoothing algorithm which generalizes existing algorithms which are used for state smoothing in hidden-Markov processes.

1.2 Speech enhancement

The problem of spectral enhancement of noisy speech signals from a single microphone has attracted considerable research effort for over thirty years, and is still an active research area, e.g., [28–40]. This problem is often formulated as estimation of speech spectral components from a degraded signal which consists of statistically independent additive noise. A variety of different approaches for spectral enhancement of noisy speech signals have been introduced over the years. One of the earlier methods is the spectral subtraction [29,30]. Accordingly, an estimate of the clean signal is obtained by subtracting an estimate of the power spectral density (PSD) of the background noise from the short-term PSD of the degraded signal. The square root of the result is considered as an estimate for the spectral magnitude of the desired signal, while the phase of the noisy signal is used as

the desired phase. This method generally results in random fluctuations in the residual noise, known as musical noise, which may be annoying and disturb the perception of the enhanced speech [41]. Many variations for the spectral subtraction method have been developed during the years to cope with the musical residual noise phenomena [42–45].

Statistical methods for speech enhancement are designed to minimize the expected value of some distortion measure between the clean and estimated signals [32–34, 36, 38, 46, 47]. These approaches require presumption of reliable statistical models for the speech and noisy signals and specification of a perceptually meaningful distortion measure. Distortion measures which are of particular interest in speech enhancement applications are the squared error [32, 48], the squared error of the short-term spectral amplitude (STSA) [33] and the squared error of the log spectral amplitude (LSA) [34,38]. To enable further attenuation of the additive noise, estimation under speech presence uncertainty is often considered [32, 33, 37, 38, 41, 49, 50]. In addition, to eliminate residual noise in case of speech absence, voice activity detector (VAD) is often incorporated with the estimator output [51–56]. Subspace methods for speech enhancement attempt to decompose the vector space of the noisy signal into a signal-plus-noise subspace and a noise-only subspace [57–60]. Spectral enhancement is then performed by removing the noise subspace and estimating the speech coefficients from the signal-plus-noise subspace. Another spectral enhancement method relies on modeling the vectors of speech signal based on hidden Markov models (HMMs) [35, 61–63]. The probability distribution of the speech (and in some applications also of the noise) are estimated from long training sequences of clean samples. The speech signal is then estimated from the noisy observation based on the trained model according to some distortion criteria. HMMs are successfully applied for speech recognition applications, e.g., [64, 65]. However, for the application of speech enhancement they were not found to be sufficiently refined models [66].

1.2.1 Spectral modeling of speech signals

Spectral enhancement of speech signals often relies on the assumption that the spectral coefficients of the speech signal (as well as of the noise signal) are statistically independent, with conditional distribution which may be considered as, e.g., Gaussian [33,34,38], Super-Gaussian [47], Laplace [67], or Gamma [68]. Although the statistical independency

assumption simplifies the design of the optimal estimator under any of the assumed distributions, it does not hold in reality since consecutive spectral magnitudes are strongly correlated [24, 25, 41, 69]. In fact, in many of the above-mentioned speech enhancement algorithms, the time-frequency dependent speech spectral variances are estimated using the decision-directed approach [33, 70] which relies on the strong correlation of successive spectral variances. Let x and d denote speech signal and uncorrelated noise signal, respectively, and let $y = x + d$ denote the observed signal. Applying the STFT to the observed signal we obtain $Y_{tk} = X_{tk} + D_{tk}$ were t and k denote the time-frame and frequency-bin indices, respectively. The decision-directed estimator for the speech spectral variance is given by

$$\hat{\lambda}_{tk} = \alpha \left| \hat{X}_{t-1,k} \right|^2 + (1-\alpha) \max \left\{ \left(|Y_{tk}|^2 - \lambda_{d,tk} \right) , 0 \right\} \tag{1.1}$$

were $\hat{\lambda}_{tk}$ denotes the estimate for the speech spectral variance and $\lambda_{d,tk}$ denotes the spectral variance of the noise spectral coefficient. The parameter α $(0 \leq \alpha \leq 1)$ is a weighting factor (typically chosen close to one) that controls the tradeoff between noise reduction and transient distortion brought into the signal [70]. A larger value of α results in a greater reduction of the musical noise phenomena, but at the expense of attenuated speech onsets and audible modifications of transient components. It is worth noting, that the noise spectral variance needs also to be estimated. This can be practically obtained during speech absence intervals or from the noisy observations by using the minima controlled recursive averaging algorithm [38, 71] or the minimum statistics approach [72].

A relaxed statistical model for speech signals was proposed in [69]. This model is based on the assumptions that speech spectral phases are iid random variables, and the speech spectral component X_{tk} is a zero-mean complex Gaussian random variable with iid real and imaginary components. The sequence of speech spectral variances $\{\lambda_{tk} \,|\, t = 0, 1, ...\}$ is a random process and each spectral variance is correlated with the sequence of the spectral magnitudes. However, given a specific spectral variance λ_{tk}, the spectral magnitude at the same time-frequency bin is statistically independent with other spectral magnitudes. This model firstly treated both the sequences of the spectral coefficients $\{X_{tk}\}$ and of the spectral variances $\{\lambda_{tk}\}$ at a specific frequency-bin as random processes.

Recently, it was proposed to model speech spectral coefficients using GARCH model [23–25]. This model explicitly parameterizes the time-varying volatility (conditional vari-

ance) at a specific frequency-bin index, by using both recent conditional variances and recent squared values of the spectral coefficients. It was shown that under perfect detection for the speech spectral coefficients, improved performance is achieved by using the GARCH model compared to using the decision-directed approach. In this dissertation, we introduce a GARCH model with Markov regimes for speech spectral modeling. This model exploits the advantages of both the conditional heteroscedasticity structure of GARCH models and the time-varying characteristics of hidden Markov chains. The model parameters are allowed to change in time according to the state of a hidden Markov chain and may be ascribed to switching between speech phonemes or different speakers. We develop model based algorithms, which are shown to be useful for speech enhancement applications, speech dereverberation, and acoustic source separation.

1.3 Other speech processing applications

While the classical problem of speech enhancement was extensively studied during recent decades, the applications of hands-free communication and digital storing of audio signals raised interesting problems such as speech dereverberation, nonstationary noise reduction, and acoustic blind source separation.

Speech signals that are received by a distant microphone from the speech source usually contain reverberation. The sound wave produced by the speaker is propagated outward from the source. The wavefronts reflect off the walls and other objects and superimposed at the microphone. These reflections can degrade the fidelity and intelligibility of speech signals and the performance of automatic speech recognition systems [73]. The received reverberated sound generally consists of a direct sound, reflections that arrive shortly after the direct sound (commonly called early reverberation), and reflections that arrive after the early reverberation (or late reverberation) [73–75]. While the early reflections mainly contribute to spectral coloration and may even improve the intelligibility of the speech sound, the late reverberation changes the waveform's temporal envelope as decaying 'tails' are added at sound offsets. This may cause *distant* and *echo-ey* sound quality [73, 76]. Algorithms for reverberation reduction (or dereverberation) can be divided to two classes. Algorithms in the first class are based on an estimation of the acoustic impulse response

(AIR). The desired signal is, in that case, estimated by deconvolution methods, e.g., [77–79]. Two drawbacks of these algorithms are, however, that they have shown to be sensitive to small changes in the AIR, and in some cases the order of the AIR needs to be known [80, 81]. Methods in the second category try to suppress reverberation without estimating the AIR, e.g., [74, 82–84]. The spectral coefficients of the desired signal are estimated using some statistical model, while trying to suppress the undesired reverberation. We develop a dual-microphone speech dereverberation algorithm for noisy environments, which is aimed at suppressing late reverberation and background noise. The spectral variance of the late reverberation is obtained with adaptively-estimated direct path compensation, and a MS-GARCH model is used to estimate the spectral variance of the desired signal, which includes the direct sound and early reverberation.

Blind 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 [85, 86], separation of different musical sources (e.g., different musical instruments) [85, 87, 88], separation of speech or singing voice from background music [89–92], and signal enhancement in nonstationary noise environments [35, 62, 93–95]. In case the signals are received by multiple microphones, spatial filtering may be employed as well as mutual statistical information between the received signals, e.g., [96–102]. However, if several sources are recorded by a single microphone, some *a priori* information is necessary to enable a reasonable separation performance.

In [94, 95] speech and nonstationary noise signals are assumed to evolve as autoregressive (AR) processes, while the *a priori* statistical information (codebooks) is obtained using a training phase and includes several sets of linear prediction coefficients. In [87, 88, 90] the acoustic signals are modeled by Gaussian mixture models (GMMs). In [35, 62] the acoustic signals are modeled by hidden Markov models (HMMs) with AR sub-sources. The assumed statistical models provide *a priori* information about the distinct signals, which together with training sequences of signals and appropriately extracted codebooks enable source separation from signal mixtures. GMM and AR-based codebooks are generally insufficient for representing statistically rich signals such as speech signals [92], since each state specifies a predetermined probability density function (pdf), or a mixture of pdf's. We develop a new algorithm for single-sensor audio source separation of speech

and music signals, which is based on MS-GARCH modeling of the speech signals and GMM for the music signals. Since in MS-GARCH model the active state only specifies the evolution of the spectral variances along time, the corresponding statistical model can take almost any values for the spectral variances. The separation of the speech from the music signal is obtained by a classification and estimation approach, which enables to control the trade-off between residual interference and signal distortion.

1.4 Detection and estimation of speech signals

In many signal processing applications as well as communication applications, the signal to be estimated is not surely present in the available noisy observation. Therefore, as specified in Section 1.2, algorithms often try to estimate the signal under uncertainty (i.e., using some *a priori* probability for the existence of the signal) [32, 33, 38, 41], or alternatively, apply an independent detector for signal presence [51–56]. This detector may be designed based on the noisy observation, or, on the estimated signal. The spectral coefficients of the speech signal are generally sparse in the STFT domain in the sense that speech is present only in some of the frames, and in each frame only some of the frequency-bins contain the significant part of the signal energy. Therefore, both signal estimation and detection are generally required while processing noisy speech signals [38, 41, 103]. However, existing algorithms often focus on estimating the spectral coefficients rather than detecting their existence. The spectral-subtraction algorithm [29, 30] contains an elementary detector for speech activity in the time-frequency domain, but it generates musical noise caused by falsely detecting noise peaks as bins that contain speech, which are randomly scattered in the STFT domain. Subspace approaches for speech enhancement [57, 59, 60, 104] decompose the vector of the noisy signal into a signal-plus-noise subspace and a noise subspace, and the speech spectral coefficients are estimated after removing the noise subspace. Accordingly, these algorithms are aimed at detecting the speech coefficients and subsequently estimating their values. McAulay and Malpass [32] were the first to propose a speech spectral estimator under a two-state model. They derived a maximum likelihood (ML) estimator for the speech spectral amplitude under speech-presence uncertainty. Ephraim and Malah followed this approach of signal estimation

under speech presence uncertainty and derived an estimator which minimizes the mean-square error (MSE) of the short-term spectral amplitude (STSA) [33]. In [49], speech presence probability is evaluated to improve the minimum MSE (MMSE) of the LSA estimator, and in [38] a further improvement of the MMSE-LSA estimator is achieved based on a two-state model.

A reliable detector for speech activity in the time-frequency domain is of major importance, not only to improve noise reduction, but also for noise estimation, speech coding and speech recognition. Many VAD algorithms are designed on a frame-by-frame basis, e.g., [51–56], or the activity of speech is detected in each time-frequency bin, e.g., [38,103]. The detection of speech presence by using a microphone array has recently received much attention, e.g., [105–108]. Spriet *et al.* [109] analyzed the impact of speech detection errors on the noise reduction performance of multichannel systems and showed that a reliable speech detector is crucial to achieve a potentially better speech enhancement performance.

Approaches for the design of coupled operations of signal detection and estimation have been proposed for some communication applications, e.g., [110–113]. In [114] a method for optimal simultaneous classification and estimation has been proposed by minimizing the erroneous classification probability in the worst case under a false alarm constraint. Middleton *et al.* [115, 116] were the first to propose simultaneous signal detection and estimation within the framework of statistical decision theory. They proposed some dual schemes for the two operations while considering several coupling methods between the two operations. The detector is generally optimized with the knowledge of the specific structure of the estimator, and the estimator is optimized in the sense of minimizing a Bayes risk associated with the combined operations.

In this research, we reformulate the speech enhancement problem as a joint problem of speech activity detection and spectral estimation. The problem formulation assumes sparsity of the speech spectral coefficients and it introduces coupled operations of detection and estimation using a combined Bayes risk. The Bayes risk incorporates both the cost of estimation errors and the cost of fault detection, whether it is a missed-detection of speech components or a false detection. While considering the problem of single-channel audio source separation, multi-hypotheses are used for both signals. In that case we incorporate simultaneous classification and estimation scheme for the separation. In both

cases, the combined cost results in cost parameters that enable to control the trade-off between signal distortion, caused by missed detection of speech components, and residual noise resulting from false-detection.

1.5 Thesis structure

This thesis is organized as follows. Chapter 2 briefly outlines the basic theories and methods which were used during this research. The original contribution of this research starts in Chapter 3.

In Chapter 3, we develop a comprehensive approach for stationarity analysis of MS-GARCH models. We consider the general case of m-state Markov chains and (p, q)-order GARCH processes, and specify the unconditional variance of the process using the expectation of the regime dependent conditional variances. No history knowledge of the process is assumed, except for the model parameters. The expectation of the conditional variance at a given regime is then recursively constructed from the conditional expectation of both previous conditional and unconditional variances. Consequently, we obtain a complete recursion for the expected vector of state dependent conditional variances. The recursive vector form is constructed by means of a representative matrix which is built from the model parameters. We show that constraining the largest absolute eigenvalue of the representative matrix to be less than one is necessary and sufficient for the convergence of the unconditional variance, and therefore, for the asymptotic stationarity of the process. We derive stationarity conditions for the general formulation of the two variants of MS-GARCH models introduced by Klaassen [7] and Haas *et al.* [8]. We show that our results reduce in some degenerated cases to the stationarity conditions developed by Bollerslev [1], and by Klaassen and Haas *et al.*. Furthermore, we show that the stationarity conditions developed by Klaassen are not only necessary but also sufficient for asymptotic stationarity of his model.

In Chapter 4, we introduce a Markov-switching time-frequency GARCH (MSTF-GARCH) model for speech signals in the STFT domain. The MSTF-GARCH model exploits the advantages of both the conditional heteroscedasticity structure of GARCH models and the time-varying characteristics of hidden Markov chains. The expansion

coefficients are considered as nonstationary random signals in the time-frequency domain and modeled as multivariate complex GARCH processes with Markov-switching regimes. A corresponding recursive algorithm is developed for signal restoration in a noisy environment. The conditional variance is estimated by iterating propagation and update steps with regime conditional probabilities. The model parameters are estimated from a training data set prior to the signal restoration using ML approach, and the number of states is assumed to be known. We show that the derivation in [117] of bounds on the MSE of a composite source signal estimation is applicable for obtaining an upper bound on the MSE of a single step MSTF-GARCH estimation. Experimental results demonstrate the improved performance of the proposed algorithm for restoration of MSTF-GARCH process compared to using an estimator which assumes a stationary process and compared to using an estimator which assumes a smaller number of regimes than the process actually has. Furthermore, it is demonstrated that the squared absolute values of speech coefficients in the STFT domain are better evaluated by using the MSTF-GARCH model than by using the decision-directed approach.

In Appendix 4.A, we present an application of the MSTF-GARCH model to speech enhancement. We employ the MSTF-GARCH model by assuming different Markov chains in distinct frequency subbands with identical state transition probabilities. The GARCH parameters are state dependent and frequency variant. We define an additional state for the case where speech coefficients are absent (or below a certain threshold level) and introduce parameter estimation method which is computationally more efficient than the traditional ML approach. Furthermore, the probability of the speech absence state can be used as a soft voice activity detector which is naturally generated in the reconstruction algorithm. Experimental results demonstrate improved noise reduction performance while preserving weak components of the speech signal.

Chapter 5 addresses the problem of state smoothing in MS-GARCH processes in noisy environments. The dependency of the conditional variance on past observations and past active regimes are taken into consideration as we generalize both the forward-backward recursions [118] and the stable backward recursion [119, 120]. We derive two recursive steps for the evaluation of conditional densities of future observations. The first step is an upward recursion which manipulates the future observations for the evaluation of

their conditional densities, corresponding to all possible future paths. The second step is a backward recursion which integrates over these paths to evaluate the future densities required for the noncausal state probability. The computational complexity of the generalized recursions grows exponentially with the number of future observations employed for the fixed-lag smoothing. However, experimental results demonstrate that the significant part of the improvement in performance, compared to using causal estimation, is achieved by considering a few future observations.

In Chapter 6, we propose a new formulation for the speech enhancement problem based on simultaneous operations of detection and estimations. A detector for the speech coefficients is combined with an estimator, which jointly minimizes a cost function that takes into account both estimation and detection errors. Under speech-presence, the cost is proportional to a quadratic spectral amplitude (QSA) error [33], while under speech-absence, the distortion depends on a certain attenuation factor [29, 38, 70]. We derive a combined detector and estimator with cost parameters that enable to control the trade-off between speech distortion, caused by missed detection of speech components, and residual musical noise resulting from false-detection. The combined solution generalizes the well-known STSA algorithm, which involves merely estimation under signal presence uncertainty. In addition, we propose a modification of the decision-directed *a priori* SNR estimator, which is suitable for transient-noise environments. Experimental results show that the simultaneous detection and estimation yields better noise reduction than the STSA algorithm while not degrading the speech signal. The advantage of using a suitable indicator for transient noise is demonstrated in a nonstationary noise environment, where the proposed algorithm facilitates suppression of transient noise with a controlled level of speech distortion.

Appendix 6.B introduces a closely related application of removal of transient noise using a practical detector for the presence of transient noise. We formulate a speech enhancement problem under multiple hypotheses, assuming some indicator or detector for the presence of noise transients in the STFT domain is available. Cost parameters control the trade-off between speech distortion and residual transient noise. We derive an optimal signal estimator that employs the available detector and show that the resulting estimator generalizes the optimally-modified log-spectral amplitude (OM-LSA) estimator [38].

Experimental results demonstrate the improved performance obtained by the proposed algorithm, compared to using the OM-LSA.

Chapter 7 deals with the problem of single-channel audio source separation. A novel approach is proposed for single-channel blind source separation of speech and music signals. This approach includes a new codebook for speech signals, as well as a new separation algorithm which relies on a simultaneous classification and estimation procedure. The codebook is based on GARCH modeling of speech signals and a GMM for music signals. Two methods are proposed for classification and estimation. One is based on simultaneous operations of classification and estimation which jointly minimize a combined Bayes risk. The second method employs a given (non-optimal) 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 with several states of parameters enables smooth (diagonal) 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 enables the user to control the trade-off between the distortion of the desired signal caused by missed detection, and the amount of the residual signal resulting from false detection.

In Chapter 8, we consider the problem of speech dereverberation using MS-GARCH modeling. We develop an improved dual-microphone speech dereverberation algorithm which relies on a MS-GARCH modeling of the desired early speech component, which consists of the direct sound and early reverberation. The model is applied to distinctive frequency subbands and specifies the volatility clustering of successive spectral coefficients, while a speech-absence state is used for evaluating the speech presence probability. Furthermore, an adaptive approach is developed to estimate the parameter for the direct path compensation (DPC) directly from the observed signals. Experimental results show that using the MS-GARCH modeling rather than the decision-directed approach, improved results can be obtained. Furthermore, by using the proposed algorithm, the performance obtained with blindly estimated DPC parameter is comparable to that obtained with an optimal DPC parameter that is calculated from the actual AIR, which is

unknown in practice.

Chapter 9 summarizes the main contributions of this dissertation and presents some future research directions.

1.6 List of publications

The chapters of this thesis are based on the following publications:

Chapter 3 is based on:

1. A. Abramson and I. Cohen, *On the stationarity of Markov switching GARCH processes*, Econometric Theory, vol. 23, no. 3, pp.485-500, 2007.

Chapter 4 is based on:

2. A. Abramson and I. Cohen, *Recursive supervised estimation of a Markov-switching GARCH process in the short-time Fourier transform domain*, IEEE Trans. on Signal Processing, vol. 55, no. 7, pp. 3227-3238, July 2007.

3. A. Abramson and I. Cohen, *Asymptotic stationarity of Markov-switching time-frequency GARCH processes*, in Proc. 30th IEEE Internat. Conf. Acoust. Speech Signal Processing., ICASSP-06, Toulouse, France May 2006, pp. III 452-455.

Appendix 4.A is based on:

4. A. Abramson and I. Cohen, *Markov-switching GARCH model and application to speech enhancement in subbands*, in Proc. 10th Internat. Workshop on Acous. Echo and Noise Control, IWAENC-2006, Paris, France September 2006. (Best student paper).

Chapter 5 is based on:

5. A. Abramson and I. Cohen, *State smoothing in Markov-switching time-frequency GARCH models*, IEEE Signal Processing Letters, vol. 13, no. 6, pp. 377-380, June 2006.

Chapter 6 is based on:

6. A. Abramson and I. Cohen, *Simultaneous detection and estimation approach for speech enhancement*, IEEE Trans. Audio, Speech, and Language Processing, vol 15, no. 8, pp. 2348-2359, Nov. 2007.

Appendix 6.B is based on:

7. A. Abramson and I. Cohen, *Enhancement of speech signals under multiple hypotheses using an indicator for transient noise presence*, Proc. 31th IEEE Internat. Conf. Acoust. Speech Signal Processing., ICASSP-07, pp. IV 533-536, Honolulu, Hawaii Apr. 2007. (Best student paper finalist).

Chapter 7 is based on:

8. A. Abramson and I. Cohen, *Single-sensor audio source separation using classification and estimation approach and GARCH modeling*, submitted to IEEE Trans. Audio, Speech, and Language Processing.

and Chapter 8 is based on:

9. A. Abramson, Emanuël A. P. Habets, S. Gannot, and I. Cohen, *Dual-microphone speech dereverberation using GARCH modeling*, to appear in Proc. 32th IEEE Internat. Conf. Acoust. Speech Signal Processing., ICASSP-08, Las-Vegas, Apr. 2008.

Chapter 2

Research Methods

In this chapter, we briefly review research methods which were useful during this research. We start by introducing the GARCH model formulation. We then continue by introducing a specific formulation for MS-GARCH model with its known conditions for asymptotic stationarity. The GARCH modeling approach for speech spectral coefficients is briefly reviewed with the variance estimation algorithm, as well as spectral enhancement approach. Finally, we briefly review existing methods for single-sensor blind source separation and speech dereverberation by using spectral enhancement.

2.1 GARCH Models and stationarity analysis

A linear (p, q)-order GARCH model is defined as follows. Let ε_t denote a real-valued discrete-time stochastic process, and let ψ_t denote the information set (σ-field) of all information through time t. The GARCH(p, q) process is then given by [1]

$$\varepsilon_t = \sigma_t \, v_t \tag{2.1}$$

where $\{v_t\}$ are iid random variables with zero mean, unit variance, and some predetermined probability density. The conditional variance of the process, $\sigma_t^2 = E\{\varepsilon_t^2 \,|\, \psi_{t-1}\}$, evolves as

$$\sigma_t^2 = \xi + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \tag{2.2}$$

where

$$\begin{aligned} p &\geq 0, \quad q > 0\,, \\ \xi &> 0, \quad \alpha_i \geq 0, \quad i = 1, ..., q, \\ \beta_j &\geq 0, \quad j = 1, ..., p. \end{aligned}$$

For $p = 0$ the process reduces to the ARCH(q) process [2], and for $q = p = 0$ the process ε_t is simply a white noise. The nonnegativity of the parameters $\alpha_i,\ i = 1, ..., q$ and $\beta_j,\ j = 1, ..., p$, together with $\xi > 0$, are sufficient to ensure a positive conditional variance, σ_t^2. However, since the conditional variance is a time varying random process, it is net guarantied in general that the process would be finite.

Theorem 2.1. *[1] The GARCH(p, q) process as defined in (2.1) and (2.2) is (asymptotically) wide-sense stationary with* $E(\varepsilon_t) = 0$, $\lim_{t\to\infty} Var(\varepsilon_t) = \xi / \left(\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j\right)$ *and* $Cov(\varepsilon_t, \varepsilon_\tau) = 0$ *for* $t \neq \tau$ *if and only if* $\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1$.

The proof can be found in [1, 5].

Theorem 2.1 gives important constraint on the model parameters such that the second-order moment of the GARCH process would be finite. It also shows that under this condition, the process is asymptotically wide-sense stationary.

2.1.1 Markov-switching GARCH model

Let $S_t \in \{1, ..., m\}$ denote the (unobserved) regime at a discrete time t and let s_t be a realization of S_t, assuming that $\{S_t\}$ is a first-order stationary Markov chain with transition probabilities $a_{ij} \triangleq p(S_t = j \,|\, S_{t-1} = i)$, a transition probabilities matrix A, $\{A\}_{ij} = a_{ij}$, and stationary probabilities $\boldsymbol{\pi} = [\pi_1, \pi_2, ..., \pi_m]'$, $\pi_i \triangleq p\,(S_t = i)$, where $'$ denotes the transpose operation.

Incorporating GARCH models with a hidden Markov chain, where each state of the chain (regime) allows a different GARCH behavior and thus a different volatility structure, extends the dynamic formulation of the model and potentially enables improved forecasts of the volatility [6–11]. Unfortunately, the volatility of a GARCH process with switching-regimes depends on the entire history of the process, including the regime path, which makes the derivation of a volatility estimator impractical. Many different variants

have been formulated for Markov-switching GARCH models. One popular formulation in econometrics is the model proposed by Klaassen [7] as a modification of Gray's model [6]. These models integrate out the unobserved regime path so that the conditional variance can be constructed from previous observations only. Accordingly, given the Markovian active state $s_t \in \{1, 2\}$, the conditional variance follows [7]:

$$\sigma^2_{t,s_t} = \xi_{s_t} + \alpha_{s_t}\varepsilon^2_{t-1} + \beta_{s_t} E\left\{\sigma^2_{t-1,S_{t-1}} \mid s_t, \psi_{t-1}\right\} . \tag{2.3}$$

As can be seen from (2.3), this model formulation originally assumes a degenerated model with only two-state Markov chain with GARCH$(1, 1)$ in each state. Define a 2×2 matrix C with elements $c_{ij} = p(S_{t-1} = j \mid S_t = i)(\alpha_i + \beta_i)$. Then, we have the following theorem:

Theorem 2.2. *[7] Necessary conditions for asymptotic wide-sense stationarity of the process defined in (2.1) and (2.3) are* $c_{11}, c_{22} < 1$ *and* $\det(I - C) > 0$. *The asymptotic conditional variances are then given by:*

$$\begin{bmatrix} \sigma^2_1 \\ \sigma^2_1 \end{bmatrix} = (I - C)^{-1} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} . \tag{2.4}$$

Proof. The unconditional variance under s_t can be obtained by taking expectation which is conditioned on the active state

$$\begin{aligned} \sigma^2_{t,s_t} &= \xi_{s_t} + \alpha_{s_t} E\left[\varepsilon^2_{t-1} \mid s_t\right] + \beta_{s_t} E\left[\sigma^2_{t-1,S_{t-1}} \mid s_t\right] \\ &= \xi_{s_t} + \alpha_{s_t} E\left\{E\left[\varepsilon^2_{t-1} \mid s_{t-1}, s_t\right] \mid s_t\right\} + \beta_{s_t} E\left\{E\left[\sigma^2_{t-1,S_{t-1}} \mid s_{t-1}, s_t\right] \mid s_t\right\} \\ &= \xi_{s_t} + \alpha_{s_t} E\left\{\sigma^2_{t-1,s_{t-1}} \mid s_t\right\} + \beta_{s_t} E\left\{\sigma^2_{t-1,s_{t-1}} \mid s_t\right\} \\ &= \xi_{s_t} + (\alpha_{s_t} + \beta_{s_t}) E\left\{\sigma^2_{t-1,s_{t-1}} \mid s_t\right\} . \end{aligned} \tag{2.5}$$

Next, assume that $\sigma^2_{t,1}$ and $\sigma^2_{t,2}$ are time invariant, and denote them by σ^2_1 and σ^2_2, respectively. Then

$$\begin{bmatrix} \sigma^2_1 \\ \sigma^2_1 \end{bmatrix} = \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + C \begin{bmatrix} \sigma^2_1 \\ \sigma^2_1 \end{bmatrix} \tag{2.6}$$

where $c_{ij} = p\left(S_{t-1} = j \mid S_t = i\right)(\alpha_i + \beta_i)$ and

$$p\left(s_{t-1} \mid s_t\right) = \frac{p\left(s_{t-1} \mid s_t\right) p\left(s_{t-1}\right)}{p\left(S_{t-1} = 1 \mid s_t\right) p\left(S_{t-1} = 1\right) + p\left(S_{t-1} = 2 \mid s_t\right) p\left(S_{t-1} = 2\right)} . \tag{2.7}$$

Under the assumption that σ^2_1 and σ^2_2 exist, $(I - C)$ is invertible and we have (2.4).

The necessary conditions for the existence of the unconditional variances are derived as follows. Since we have sufficient conditions for the positivity of the conditional variances (i.e., $\xi_s > 0$ and $\alpha_s, \beta_s \geq 0$ for $s = 1, 2$), all elements of $(I - C)^{-1}$ must be non negative and $(I - C)^{-1}$ must not have a zero raw. Since $c_{12} = \alpha_1 + \beta_1 - c_{11}$ and $c_{21} = \alpha_2 + \beta_2 - c_{22}$ we can write

$$(I - C)^{-1} = \frac{1}{\det(I - C)} \begin{bmatrix} 1 - c_{22} & \alpha_1 + \beta_1 - c_{11} \\ \alpha_2 + \beta_2 - c_{22} & 1 - c_{11} \end{bmatrix}. \tag{2.8}$$

Since

$$\alpha_i + \beta_i - c_{ii} = (\alpha_i + \beta_i)\left[1 - p\left(S_{t-1} = i \,|\, S_t = i\right)\right] \geq 0 \tag{2.9}$$

the nonnegativity of (2.8) implies that $\det(I - C) > 0$, so that $1 - c_{11} \geq 0$ and $1 - c_{22} \geq 0$. But c_{11} and c_{22} must not equal one, otherwise, $\det(I - C) \leq 0$. Therefore, necessary conditions for the existence of stationary variances are $c_{11}, c_{22} < 1$ and $\det(I - C) > 0$. □

In Chapter 3 we derive conditions which are both necessary and sufficient for asymptotic stationarity, considering any finite-state Markov chain and any (p, q)-order of GARCH model in each state.

2.2 Time-frequency GARCH model and spectral speech enhancement

2.2.1 Time-frequency GARCH model

Recall x and d represent speech and uncorrelated additive noise signals, and $y = x + d$ represents the observed signal. Applying the STFT to the observed signal we have in the time-frequency domain

$$Y_{tk} = X_{tk} + D_{tk}\,. \tag{2.10}$$

Let $\mathcal{X}^\tau = \{X_{tk} \,|\, t = 0, 1, ..., \tau,\ k = 0, ..., K - 1\}$ represent the set of clean spectral coefficients up to frame τ. Let $\mathcal{H}_1^{tk}$ and $\mathcal{H}_0^{tk}$ denote hypotheses for speech present and absence, respectively, in the time-frequency bin (t, k), and let

$$\lambda_{tk|\tau} \triangleq E\left\{|X_{tk}|^2 \,|\, \mathcal{H}_1^{tk}, \mathcal{X}^\tau\right\} \tag{2.11}$$

denote the *conditional* variance of X_{tk} under the hypothesis that speech is present in the time-frequency bin (t,k), given the clean spectral coefficients up to time-frame τ. The time-frequency GARCH (TF-GARCH) relies on the following assumptions [23, 25]:

1. Given $\{\lambda_{tk}\}$ and the state of speech presence in each time-frequency bin ($\mathcal{H}_1^{tk}$ or $\mathcal{H}_0^{tk}$), the speech spectral coefficients $\{X_{tk}\}$ are generated by

$$X_{tk} = \sqrt{\lambda_{tk}} V_{tk}\,, \tag{2.12}$$

   where $\left\{V_{tk} \,|\, \mathcal{H}_0^{tk}\right\}$ are identically zero, and $\left\{V_{tk} \,|\, \mathcal{H}_1^{tk}\right\}$ are statistically independent complex random variable with zero mean, unit variance, and independent and identically distributed (iid) real and imaginary parts:

$$\begin{aligned} &\mathcal{H}_1^{tk} : E\left\{V_{tk}\right\} = 0,\ E\left\{|V_{tk}|^2\right\} = 1\,, \\ &\mathcal{H}_1^{tk} : V_{tk} = 0\,. \end{aligned} \tag{2.13}$$

2. Under $\mathcal{H}_1^{tk}$, the real and imaginary parts of V_{tk} are iid with Gaussian probability density.

3. The conditional variance $\lambda_{tk|t-1}$, referred to as the *one-frame-ahead conditional variance*, is a random process which evolves as a GARCH$(1,1)$ process:

$$\lambda_{tk|t-1} = \lambda_{\min} + \alpha\,|X_{t-1,k}|^2 + \beta\left(\lambda_{t-1,k|t-2} - \lambda_{\min}\right) \tag{2.14}$$

   where

$$\lambda_{\min} > 0,\ \alpha \geq 0,\ \beta \geq 0,\ \alpha + \beta < 1\,. \tag{2.15}$$

4. The noise spectral coefficients $\{D_{tk}\}$ are zero-mean statistically independent Gaussian random variables, with iid real and imaginary parts.

The first assumption implies that the speech spectral coefficients $\left\{X_{tk} \,|\, \mathcal{H}_1^{tk}\right\}$ are conditionally zero-mean statistically independent random variables given their variances $\{\lambda_{tk}\}$. However, the GARCH formulation parameterizes the correlation between successive conditional variances at the same frequency-bin index.

2.2.2 Variance estimation

Since the conditional variance in the TF-GARCH depends on the entire history of the process, while observing noisy signal, the conditional variances can not be reconstructed and need to be estimated. The estimation of the spectral variance from the noisy observations is estimated by using two steps. First, the conditional variance estimate $\hat{\lambda}_{tk|t-1}$ is being updated one frame ahead in time by using the additional information Y_{tk}, than, for the next frame the conditional variance is updated based on the model formulation. Specifically, an estimate for $\lambda_{tk|t}$ is obtained by calculating its conditional mean under $\mathcal{H}_1^{tk}$ given Y_{tk} and $\hat{\lambda}_{tk|t-1}$. By definition $\lambda_{tk|t} = |X_{tk}|^2$. Hence, the update step is obtained by [24, 69]:

$$\begin{aligned}\hat{\lambda}_{tk|k} &= E\left\{|X_{tk}|^2 \,\middle|\, \mathcal{H}_1^{tk}, \hat{\lambda}_{tk|t-1}, Y_{tk}\right\} \\ &= \frac{\hat{\xi}_{tk|t-1}}{1+\hat{\xi}_{tk|t-1}}\left(\lambda_{d,tk} + \frac{\hat{\xi}_{tk|t-1}}{1+\hat{\xi}_{tk|t-1}}\right)|Y_{tk}|^2 \end{aligned} \tag{2.16}$$

where $\lambda_{d,tk} = E\{|D_{tk}|^2\}$ and $\hat{\xi}_{tk|t-1} \triangleq \hat{\lambda}_{tk|t-1}/\lambda_{d,tk}$ is the *a priori* SNR. Substituting (2.16) into the model formulation (2.14) we have the propagation step:

$$\begin{aligned}\hat{\lambda}_{tk|t-1} &= E\left\{\lambda_{tk|t-1} \,|\, \mathcal{H}_1^{t-1,k}, \hat{\lambda}_{t-1,k|t-2}, Y_{t-1,k}\right\} \\ &= \lambda_{\min} + \alpha E\left\{|X_{t-1,k}|^2 \,|\, \mathcal{H}_1^{t-1,k}, \hat{\lambda}_{t-1,k|t-2}, Y_{t-1,k}\right\} + \beta\left(\hat{\lambda}_{t-1,k|t-2} - \lambda_{\min}\right) \\ &= \lambda_{\min} + \alpha\hat{\lambda}_{t-1,k|t-1} + \beta\left(\hat{\lambda}_{t-1,k|t-2} - \lambda_{\min}\right). \end{aligned} \tag{2.17}$$

This two-step estimation method allows estimation of the conditional variances from the noisy coefficients. It is important to note that the model parameters are generally estimated from a training set using maximum likelihood (ML) approach [5].

2.2.3 Spectral enhancement

Having an estimate for the spectral variances of the speech spectral coefficients, $\hat{\lambda}_{tk}$, an estimator for the coefficients X_{tk} is obtained by minimizing the expected distortion given $\hat{\lambda}_{tk}$, $\lambda_{d,tk}$, Y_{tk} and the a posteriori speech presence probability $q_{tk} \triangleq p\left(\mathcal{H}_1^{tk} \,|\, Y_{tk}\right)$ [41]:

$$\min_{\hat{X}_{tk}} E\left\{d\left(X_{tk}, \hat{X}_{tk}\right) \,|\, q_{tk}, \hat{\lambda}_{tk}, \lambda_{d,tk}, Y_{tk}\right\}. \tag{2.18}$$

In particular, restricting ourselves to a squared error distortion measure of the form

$$d\left(X_{tk},\hat{X}_{tk}\right)=\left|g\left(\hat{X}_{tk}\right)-\tilde{g}\left(X_{tk}\right)\right|^{2} \tag{2.19}$$

where $g\left(X\right)$ and $\tilde{g}\left(X\right)$ are specific functions which determine the fidelity criteria, the optimal estimator is calculated from

$$\begin{aligned} g\left(\hat{X}_{tk}\right) &= E\left\{\tilde{g}\left(X_{tk}\right)\mid q_{tk},\hat{\lambda}_{tk},\lambda_{d,tk},Y_{tk}\right\} \\ &= q_{tk}\,E\left\{\tilde{g}\left(X_{tk}\right)\mid \mathcal{H}_{1}^{tk},\hat{\lambda}_{tk},\lambda_{d,tk},Y_{tk}\right\} \\ &\quad +\left(1-q_{tk}\right)E\left\{\tilde{g}\left(X_{tk}\right)\mid \mathcal{H}_{0}^{tk},Y_{tk}\right\}. \end{aligned} \tag{2.20}$$

Fidelity criteria that are of particular interest for speech enhancement applications are the MMSE of the STSA [33] and MMSE of the LSA [34]. The MMSE-STSA estimator is derived by substituting into (2.19) the functions

$$\begin{aligned} g\left(\hat{X}_{tk}\right) &= \left|\hat{X}_{tk}\right| \\ \tilde{g}\left(X_{tk}\right) &= \begin{cases} \left|X_{tk}\right|, & \text{under } \mathcal{H}_{1}^{tk} \\ G_{\min}\left|Y_{tk}\right|, & \text{under } \mathcal{H}_{0}^{tk} \end{cases}. \end{aligned} \tag{2.21}$$

Let $\gamma_{tk}=|Y_{tk}|^{2}/\lambda_{d,tk}$ denote the *a posteriori* SNR and let $\upsilon_{tk}\triangleq\gamma_{tk}\,\hat{\xi}_{tk}/(1+\hat{\xi}_{tk})$. The resulting estimator is given by

$$\hat{X}_{tk}=\left[q_{tk}\,G_{STSA}\left(\hat{\xi}_{tk},\gamma_{tk}\right)+\left(1-q_{tk}\right)G_{\min}\right]Y_{tk} \tag{2.22}$$

where [33]

$$G_{STSA}\left(\xi,\gamma\right)\triangleq\frac{\sqrt{\pi\,\upsilon}}{2\gamma}\exp\left(-\frac{\upsilon}{2}\right)\left[\left(1+\upsilon\right)I_{0}\left(\frac{\upsilon}{2}\right)+\upsilon\,I_{1}\left(\frac{\upsilon}{2}\right)\right], \tag{2.23}$$

and $I_{\nu}\left(\cdot\right)$ denotes the modified Bessel function of order ν. The MMSE-LSA estimator is obtained by using the functions

$$\begin{aligned} g\left(\hat{X}_{tk}\right) &= \log\left|\hat{X}_{tk}\right| \\ \tilde{g}\left(X_{tk}\right) &= \begin{cases} \log\left|X_{tk}\right|, & \text{under } \mathcal{H}_{1}^{tk} \\ \log\left(G_{\min}\left|Y_{tk}\right|\right), & \text{under } \mathcal{H}_{0}^{tk}. \end{cases} \end{aligned} \tag{2.24}$$

and the resulting estimator follows

$$\hat{X}_{tk}=G_{LSA}\left(\hat{\xi}_{tk},\gamma_{tk}\right)^{q_{tk}}G_{\min}^{1-q_{tk}}\,Y_{tk} \tag{2.25}$$

where [34]

$$G_{LSA}\left(\xi,\gamma\right)=\frac{\xi}{1+\xi}\exp\left(\frac{1}{2}\int_{\upsilon}^{\infty}\frac{e^{-x}}{x}dx\right). \tag{2.26}$$

It is important to note, that both estimators (2.22) and (2.25) are insensitive to the phase estimation error, and they are combined with the phase of the noisy signal [33].

2.3 Single-channel blind source separation

Separation of a mixture of signals observed via a single sensor is an ill posed problem, and some *a priori* information is required to enable reasonable reconstructions. In [87–89], a GMM is proposed for the signals' codebook in the STFT domain, and in [94, 95] an AR model is proposed with different sets of prediction coefficients for each of the signals in the time domain. However, each set of AR coefficients, together with the excitation variance corresponds to a specific covariance matrix in the STFT domain, similarly to the GMM. Under each of these models, each framed signal is considered as generated from some specific distribution which is related to the codebook with some probability, and a frame-by-frame separation is applied.

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 $p_1\left(i\right)\triangleq p\left(q_1=i\right)$, $i=0,...,m_1$ and $p_2\left(j\right)\triangleq p\left(q_2=j\right)$, $j=0,...,m_2$, and $\sum_i p_1\left(i\right)=\sum_j p_2\left(j\right)=1$. Given that $q_1=i$ and $q_2=j$, $\mathbf{s}_1$ and $\mathbf{s}_2$ are assumed conditionally zero-mean complex-valued Gaussian random vectors with known diagonal covariance matrices, i.e. ,$\mathbf{s}_1\sim\mathcal{CN}\left(0,\Sigma_1^{(i)}\right)$ and $\mathbf{s}_2\sim\mathcal{CN}\left(0,\Sigma_2^{(j)}\right)$. For the AR model [35, 62, 94, 95], each set of prediction coefficients in the time domain corresponds to a specific covariance matrix in the STFT domain, up to scaling by the excitation variance. Assuming sufficiently long frames, these covariance matrices are considered as diagonal [95].

Based on a given codebook, it is proposed in [88] and [95] to first find the active pair of states $\{i,j\}=\{q_1=i,q_2=j\}$ using a maximum a posteriori (MAP) criterion:

$$\left\{\hat{i},\hat{j}\right\}=\arg\max_{i,j}p\left(\mathbf{x}\,|\,i,j\right)p\left(i,j\right) \tag{2.27}$$

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 mmse sense by

$$\begin{aligned} \hat{\mathbf{s}}_1 &= E\left\{\mathbf{s}_1 \,|\, \mathbf{x}, \hat{i}, \hat{j}\right\} \\ &= \Sigma_1^{(\hat{i})} \left(\Sigma_1^{(\hat{i})} + \Sigma_2^{(\hat{j})}\right)^{-1} \mathbf{x} \\ &\triangleq W_{\hat{i}\hat{j}}\, \mathbf{x} \end{aligned} \tag{2.28}$$

and similarly[1] $\hat{\mathbf{s}}_2 = W_{\hat{j}\hat{i}}\, \mathbf{x}$. Alternatively [35, 88, 90], the desired signal may be reconstructed in the mmse sense directly from

$$\begin{aligned} \hat{\mathbf{s}}_1 &= E\{\mathbf{s}_1 \,|\, \mathbf{x}\} \\ &= E_{ij}\{E[\mathbf{s}_1 \,|\, \mathbf{x}, i, j]\} \\ &= \sum_{i,j} p(i,j \,|\, \mathbf{x})\, W_{ij}\, \mathbf{x}\,. \end{aligned} \tag{2.29}$$

In case of additional uncorrelated stationary noise in the mixed signal, i.e.,

$$\mathbf{x} = \mathbf{s}_1 + \mathbf{s}_2 + \mathbf{d} \tag{2.30}$$

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.

2.4 Speech dereverberation

A generalized statistical reverberation model has been proposed in [73–75]. Accordingly, the AIR $h(n)$, can be split into two segments, $h_e(n)$ and $h_l(n)$:

$$h(n) = \begin{cases} h_e(n)\,, & 0 \le n \le T_r \\ h_l(n)\,, & n \ge T_r \\ 0\,, & \text{otherwise} \end{cases}. \tag{2.31}$$

The value T_r is chosen such that $h_e(n)$ contains the direct path, and that $h_l(n)$ contains of all later reflections. To enable modeling the energy related to the direct path, the

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

following model is used:

$$h_e(n) = \begin{cases} b_e(n)\, e^{-\delta n}, & 0 \le n \le T_r \\ 0, & \text{otherwise} \end{cases}, \tag{2.32}$$

where $b_e(t)$ is a white zero-mean Gaussian stationary noise signal and δ is linked to the reverberation time[2], T_{60}. The reverberation component $h_l(t)$ follows

$$h_l(t) = \begin{cases} b_l(t)\, e^{-\delta t}, & t \ge T_r \\ 0, & \text{otherwise} \end{cases}, \tag{2.33}$$

where $b_l(n)$ is a white zero-mean Gaussian stationary noise signal. It is assumed that $b_e(n)$ and $b_l(n)$ are uncorrelated, and the energy envelope of $h(n)$ can be expressed as

$$E_h\left\{h^2(n)\right\} = \begin{cases} \sigma_e^2 e^{-\delta n}, & 0 \le n \le T_r \\ \sigma_l^2 e^{-\delta n}, & n \ge T_r \\ 0, & \text{otherwise} \end{cases}, \tag{2.34}$$

where σ_e^2 and σ_l^2 denote the variances of $b_e(n)$ and $b_l(n)$, respectively, and generally, $\sigma_e^2 \ge \sigma_l^2$.

Considering a speech signal $x(n)$ which is propagates towards a microphone, through a room with AIR $h(n)$, the received signal can be denoted as

$$y(n) = x_e(n) + x_l(n) + d(n), \tag{2.35}$$

where $x_e(n)$ is the early speech component, $x_l(n)$ is the late reverberant signal, and $d(n)$ is an uncorrelated additive noise. Spectral enhancement approach for speech dereverberation is aimed at estimating the early speech component from the noisy observation by using a time-frequency dependent gain function. Consequently, in the STFT domain we have

$$\hat{X}_e(t,k) = G(t,k)\, Y(t,k). \tag{2.36}$$

The gain function may be obtained by means of spectral subtraction [74] or MMSE-LSA sense [75]. In any case, the spectral variances of both the early speech component, $\lambda_e(t,k) = E\left\{|X_e(t,k)|^2\right\}$, and of the late reverberant signal, $\lambda_l(t,k) = E\left\{|X_l(t,k)|^2\right\}$,

[2]The reverberation time, T_{60}, is defined as the time for the reverberation level to decay to 60 dB below the initial level.

are estimated from the observed signal based on the reverberation model. Accordingly, while evaluating the gain function, the *a priori* and *a posteriori* SNRs are given by

$$\begin{aligned} \hat{\xi}\left(tk\right) &= \frac{\hat{\lambda}_e\left(t,k\right)}{\hat{\lambda}_l\left(t,k\right)+\lambda_d\left(t,k\right)} \\ \hat{\gamma}\left(t,k\right) &= \frac{\left|Y\left(t,k\right)\right|^2}{\hat{\lambda}_l\left(t,k\right)+\lambda_d\left(t,k\right)}\,. \end{aligned} \tag{2.37}$$

Chapter 3

Stationarity Analysis of Markov-Switching GARCH Processes[1]

GARCH models with Markov-switching regimes are often used for volatility analysis of financial time series. Such models imply less persistence in the conditional variance than the standard GARCH model, and potentially provide a significant improvement in volatility forecast. Nevertheless, conditions for asymptotic wide-sense stationarity have been derived only for some degenerated models. In this chapter, we introduce a comprehensive approach for stationarity analysis of Markov-switching GARCH models, which manipulates a backward recursion of the model's second-order moment. A recursive formulation of the state-dependent conditional variances is developed and the corresponding conditions for stationarity are obtained. In particular, we derive necessary and sufficient conditions for the asymptotic wide-sense stationarity of two different variants of Markov-switching GARCH processes, and obtain expressions for their asymptotic variances in the general case of m-state Markov chains and (p, q)-order GARCH processes.

[1]This chapter is based on [121].

3.1 Introduction

Volatility analysis of financial time series is of major importance in many financial applications. The generalized autoregressive conditional heteroscedasticity (GARCH) model [1] has been applied quite extensively in the field of econometrics, both by practitioners and by researchers, and shown to be useful for the analysis and forecasting the volatility of time-varying processes such as those pertaining to financial markets. Incorporating GARCH models with a hidden Markov chain, where each state of the chain (regime) allows a different GARCH behavior and thus a different volatility structure, extends the dynamic formulation of the model and potentially enables improved forecasts of the volatility [6–11]. Unfortunately, the volatility of a GARCH process with switching-regimes depends on the entire history of the process, including the regime path, which makes the derivation of a volatility estimator impractical.

Cai [12] and Hamilton and Susmel [13] applied the idea of regime-switching parameters into ARCH specification. The conditional variance of an ARCH model depends only on past observations, and accordingly the restriction to ARCH models avoids problems of infinite path dependency. Gray [6], Klaassen [7] and Haas, Mittnik and Paolella [8], proposed different variants of Markov-switching GARCH models, which also avoid the problem of dependency on the regime's path. Gray introduced a Markov-switching GARCH model relying on the assumption that the conditional variance at any regime depends on the *expectation* of previous conditional variances, rather than their values. Accordingly, the conditional variance depends only on some finite set of past state-dependent, expected values via their conditional state probabilities, and thus can be constructed from past observations. Klaassen proposed modifying Gray's model by conditioning the expectation of previous conditional variances on all available observations and also on the current regime. A different concept of Markov-switching GARCH model has recently been introduced by Haas, Mittnik and Paolella. Accordingly, a finite state-space Markov chain is assumed to govern the ARCH parameters while the autoregressive behavior of the conditional variance is subject to the assumption that past conditional variances are in the same regime as that of the current one.

Markov-switching GARCH processes, as well as the standard GARCH process, are

nonstationary as their second-order moments change recursively over time. However, if these processes are asymptotically wide-sense stationary then their variances are guaranteed to be finite. A necessary and sufficient condition for the stationarity of a (single-regime) GARCH(p, q) process has been developed in [1]. Condition for the stationarity of a *natural*, path-dependent Markov-switching GARCH(p, q) model, has been developed in [122], and in [14] a deep analysis of the probabilistic structure of that model is derived with conditions for the existence of moments of any order. In [15–17], stationarity analysis has been derived for some mixing models of conditional heteroscedasticity, and conditions for the asymptotic stationarity of some AR and ARMA models with Markov-regimes has been derived in [18–22]. However, for the Markov-switching GARCH models described above, which avoid the dependency of the conditional variance on the chain's history, stationarity conditions are known in the literature only for some special cases. Klaassen [7] developed necessary (but not necessarily sufficient) conditions for stationarity of his model in the special case of two regimes and GARCH modeling of order $(1, 1)$. A necessary and sufficient stationarity condition has been developed by Haas, Mittnik and Paolella [8] for their Markov-switching GARCH model, but only in case of GARCH$(1, 1)$ behavior in each regime.

In this chapter, we develop a comprehensive approach for stationarity analysis of Markov-switching GARCH models, in the general case of m-state Markov chains and (p, q)-order GARCH processes. We specify the unconditional variance of the process using the expectation of the regime dependent conditional variances, and assume no history knowledge of the process except for the model parameters. The expectation of the conditional variance at a given regime is then recursively constructed from the conditional expectation of both previous conditional and unconditional variances. Consequently, we obtain a complete recursion for the expected vector of state dependent conditional variances. The recursive vector form is constructed by means of a representative matrix which is built from the model parameters. We show that constraining the largest absolute eigenvalue of the representative matrix to be less than one is necessary and sufficient for the convergence of the unconditional variance, and therefore, for the asymptotic stationarity of the process. We derive stationarity conditions for the general formulation of the two variants of Markov-switching GARCH models introduced by Klaassen and Haas *et al.*.

We show that our results reduce in some degenerated cases to the stationarity conditions developed by Bollerslev [1], Klaassen [7] and Haas *et al.* [8]. Furthermore, we show that the stationarity conditions developed by Klaassen are not only necessary but also sufficient for asymptotic stationarity of his model.

This chapter is organized as follows: In Section 3.2, we review the variants proposed by Klaassen and Haas *et al.* for Markov-switching GARCH models, and develop comprehensive necessary and sufficient conditions for asymptotic stationarity appropriate for the general formulation of the models. In Section 3.3, we derive relations between our results and previous works.

3.2 Stationarity of Markov-switching GARCH models

Let $S_t \in \{1, ..., m\}$ denote the (unobserved) regime at a discrete time t and let s_t be a realization of S_t, assuming that $\{S_t\}$ is a first-order stationary Markov chain with transition probabilities $a_{ij} \triangleq p(S_t = j \,|\, S_{t-1} = i)$, a transition probabilities matrix A, $\{A\}_{ij} = a_{ij}$, and stationary probabilities $\boldsymbol{\pi} = [\pi_1, \pi_2, ..., \pi_m]'$, $\pi_i \triangleq p\,(S_t = i)$, where $'$ denotes the transpose operation. Let $\mathcal{I}_t$ denote the observation set up to time t, and let $\{v_t\}$ be a zero-mean unit-variance random process, with independent and identically distributed elements. Given that $S_t = s_t$, a Markov-switching GARCH model of order (p, q) can be formulated as

$$\varepsilon_t = \sigma_{t,s_t} v_t \tag{3.1}$$

where the conditional variance of the process $\sigma^2_{t,s_t} = E\left\{\varepsilon_t^2 \,|\, S_t = s_t, \mathcal{I}_{t-1}\right\}$ is a function of p previous conditional variances and q previous squared observations.

Klaassen [7] and Haas, Mittnik and Paolella [8], proposed different variants of Markov-switching GARCH models. The former is a modification of Gray's model [6]. Each of these overcomes the problem of dependency on the regime's path encountered when naturally integrating the GARCH model with switching-regimes. However, conditions for these models to be asymptotically wide-sense stationary and therefore to guarantee a

finite second-order moments, are known only for some special cases. Klaassen developed necessary conditions for the stationarity of his model in the case of two-state Markov chain and GARCH of order $(1,1)$. Hass *et al.* gave a necessary and sufficient stationarity condition for their model, but this condition is restricted to a first-order GARCH model in each of the regimes (*i.e.,* $p = q = 1$). We first review these variants of Markov-switching GARCH models, which we call MSG-I and MSG-II, respectively. Then we develop necessary and sufficient conditions for their asymptotic wide-sense stationarity and derive their stationary variances.

3.2.1 MSG-I model

Gray [6] proposed to model the conditional variance of a Markov-switching GARCH model as dependent on the *expectation* of its past values over the entire set of states, rather than dependent on past states and the corresponding conditional variances. Accordingly, the state dependent conditional variance follows

$$\begin{aligned}\sigma^2_{t,s_t} &= \xi_{s_t} + \sum_{i=1}^{q} \alpha_{i,s_t} \varepsilon^2_{t-i} + \sum_{j=1}^{p} \beta_{j,s_t} E\left(\varepsilon^2_{t-j} \,|\, \mathcal{I}_{t-j-1}\right) \\ &= \xi_{s_t} + \sum_{i=1}^{q} \alpha_{i,s_t} \varepsilon^2_{t-i} + \sum_{j=1}^{p} \beta_{j,s_t} \sum_{s_{t-j}=1}^{m} p\left(S_{t-j} = s_{t-j} \,|\, \mathcal{I}_{t-j-1}\right) \sigma^2_{t-j,s_{t-j}}, \end{aligned} \tag{3.2}$$

and the following constraints

$$\xi_{s_t} > 0, \quad \alpha_{i,s_t} \geq 0, \quad \beta_{j,s_t} \geq 0, \qquad i = 1,...,q, \qquad j = 1,...,p, \qquad s_t = 1,...,m \tag{3.3}$$

are sufficient for the positivity of the conditional variance.

Gray's model integrates out the unobserved regime path so that the conditional variance can be constructed from previous observations only. As a consequence, there is no path dependency problem although GARCH effects are still allowed. Empirical analysis of modeling financial time series demonstrates that this Markov-switching GARCH model implies less persistence in the conditional variance than the standard GARCH model, and in addition, its one-step ahead volatility forecast significantly outperforms the single-regime GARCH model (see, for instance [6, 9, 11]).

Klaassen [7] proposed modifying Gray's model by replacing $p\left(S_{t-j}=s_{t-j}\,|\,\mathcal{I}_{t-j-1}\right)$ in (3.2) by $p\left(S_{t-j}=s_{t-j}\,|\,\mathcal{I}_{t-1},S_t=s_t\right)$ while evaluating σ^2_{t,s_t}. Consequently, all available observations are used, as well as the given regime in which the conditional variance is calculated. The conditional variance according to Klaassen's model (denoted here as MSG-I) is given by

$$\sigma^2_{t,s_t}=\xi_{s_t}+\sum_{i=1}^{q}\alpha_{i,s_t}\varepsilon^2_{t-i}+\sum_{j=1}^{p}\beta_{j,s_t}\sum_{s_{t-j}=1}^{m}p\left(S_{t-j}=s_{t-j}\,|\,\mathcal{I}_{t-1},S_t=s_t\right)\sigma^2_{t-j,s_{t-j}}, \qquad (3.4)$$

and the same constraints (3.3) are sufficient for the positivity of the conditional variance.

Both models integrate out the unobserved regimes for evaluating the conditional variance. However, Klaassen's model employs all the available information while Gray's model employs only part of it since it does not utilize all the available observations and the assumed regime in which the conditional variance is being calculated. Specifically, if process' regimes are highly persistent, then both the current state s_t and the previous innovation ε_{t-1} give much information about previous states and thus the conditional probability of s_{t-1} given all the observations up to time $t-1$ and the next state, is substantially different from the probability of s_{t-1} which is conditioned only on observations up to time $t-2$ [7]. In contrast to Gray, Klaassen do manipulate this information in his model while evaluating the expectation of previous conditional variances. Furthermore, the formulation (3.4) better exploits the available information, and its structure yields straightforward expressions for the multi-step ahead volatility forecasts [7, 9].

The unconditional variance of the MSG-I process, defined in (3.1) and (3.4), can be calculated as follows:

$$\begin{aligned} E\left[\varepsilon^2_t\right] &= E_{\mathcal{I}_{t-1},S_t}\left[E\left(\varepsilon^2_t\,|\,\mathcal{I}_{t-1},s_t\right)\right] \\ &= E_{S_t}\left[E_{\mathcal{I}_{t-1}}\left(\sigma^2_{t,s_t}\,|\,s_t\right)\right]=\sum_{s_t=1}^{m}\pi_{s_t}E_{\mathcal{I}_{t-1}}\left(\sigma^2_{t,s_t}\,|\,s_t\right)\,. \end{aligned} \qquad (3.5)$$

For notation simplification, we shall use $E(\cdot\,|\,s_t)$ and $p(\cdot\,|\,s_t)$ to represent $E(\cdot\,|\,S_t=s_t)$ and $p(\cdot\,|\,S_t=s_t)$, respectively, where s_t represents the regime realization at time t. Furthermore, we shall use $E_t\left(\cdot\right)$ to denote the expectation over the information up to time

t, *i.e.*, $E_{\mathcal{I}_t}(\cdot)$. The expectation of the regime dependent conditional variance follows

$$\begin{aligned} E_{t-1}\left[\sigma^2_{t,s_t} \,|\, s_t\right] &= \xi_{s_t} + \sum_{i=1}^{q} \alpha_{i,s_t} E_{t-1}\left[\varepsilon^2_{t-i} \,|\, s_t\right] \\ &\quad + \sum_{j=1}^{p} \beta_{j,s_t} \sum_{s_{t-j}=1}^{m} E_{t-1}\left[p\left(s_{t-j} \,|\, \mathcal{I}_{t-1}, s_t\right) \sigma^2_{t-j,s_{t-j}} \,|\, s_t\right], \end{aligned} \tag{3.6}$$

where the expectation over ε^2_{t-i} can be obtained by

$$\begin{aligned} E_{t-1}\left[\varepsilon^2_{t-i} \,|\, s_t\right] &= \sum_{s_{t-i}=1}^{m} \int_{\mathcal{I}_{t-1}} \varepsilon^2_{t-i}\, p\left(\mathcal{I}_{t-1} \,|s_t, s_{t-i}\right) p\left(s_{t-i} \,|\, s_t\right) d\mathcal{I}_{t-1} \\ &= \sum_{s_{t-i}=1}^{m} p\left(s_{t-i} \,|\, s_t\right) E_{t-1}\left[\varepsilon^2_{t-i} \,|\, s_{t-i}, s_t\right]. \end{aligned} \tag{3.7}$$

Note that given the current active state, the expected absolute value is independent of any future states. Therefore,

$$\begin{aligned} E_{t-1}\left[\varepsilon^2_{t-i} \,|\, s_{t-i}, s_t\right] &= E_{t-1}\left[\varepsilon^2_{t-i} \,|\, s_{t-i}\right] \\ &= \int_{\mathcal{I}_{t-i-1}} \int_{\varepsilon_{t-i}} \varepsilon^2_{t-i}\, p\left(\varepsilon_{t-i} \,|\, \mathcal{I}_{t-i-1}, s_{t-i}\right) p\left(\mathcal{I}_{t-i-1} \,|\, s_{t-i}\right) d\varepsilon_{t-i} d\mathcal{I}_{t-i-1} \\ &= E_{t-i-1}\left[E\left(\varepsilon^2_{t-i} \,|\, \mathcal{I}_{t-i-1}, s_{t-i}\right) \,|\, s_{t-i}\right] \\ &= E_{t-i-1}\left[\sigma^2_{t-i,s_{t-i}} \,|\, s_{t-i}\right]. \end{aligned} \tag{3.8}$$

Furthermore, the conditional expectation over the conditional variance in (3.6), weighted by the current state probability can be obtained by

$$\begin{aligned} E_{t-1}\left[p\left(s_{t-j} \,|\, \mathcal{I}_{t-1}, s_t\right) \sigma^2_{t-j,s_{t-j}} \,|\, s_t\right] &= \int_{\mathcal{I}_{t-1}} \sigma^2_{t-j,s_{t-j}} p\left(s_{t-j} \,|\, \mathcal{I}_{t-1}, s_t\right) p\left(\mathcal{I}_{t-1} \,|\, s_t\right) d\mathcal{I}_{t-1} \\ &= \int_{\mathcal{I}_{t-1}} \sigma^2_{t-j,s_{t-j}} p\left(\mathcal{I}_{t-1} \,|\, s_{t-j}, s_t\right) p\left(s_{t-j} \,|\, s_t\right) d\mathcal{I}_{t-1} \\ &= p\left(s_{t-j} \,|\, s_t\right) E_{t-j-1}\left[\sigma^2_{t-j,s_{t-j}} \,|\, s_{t-j}\right]. \end{aligned} \tag{3.9}$$

Consequently, the expectation of the conditional variance at a given regime s_t can be recursively constructed, according to the model definitions, from both expectation of previous conditional variances and expected squared values given the current regime s_t. Let $r = \max\{p, q\}$ and define $\alpha_{i,s} \triangleq 0$ for all $i > q$ and $\beta_{i,s} \triangleq 0$ for all $i > p$. Then, by substituting (3.7), (3.8) and (3.9) into (3.6) we obtain

$$E_{t-1}\left[\sigma^2_{t,s_t} \,|\, s_t\right] = \xi_{s_t} + \sum_{i=1}^{r} \left(\alpha_{i,s_t} + \beta_{i,s_t}\right) \sum_{s_{t-i}=1}^{m} p\left(s_{t-i} \,|\, s_t\right) E_{t-i-1}\left[\sigma^2_{t-i,s_{t-i}} \,|\, s_{t-i}\right], \tag{3.10}$$

and applying Bayes' rule we have

$$p\left(s_{t-i} \mid s_t\right) = \frac{\pi_{s_{t-i}}}{\pi_{s_t}} p\left(s_t \mid s_{t-i}\right) = \frac{\pi_{s_{t-i}}}{\pi_{s_t}} \left\{A^i\right\}_{s_{t-i},s_t} . \tag{3.11}$$

The expected state dependent conditional variance (3.10) is recursively generated from a weighted sum of its previous expected values through their conditioned probabilities and the model parameters. Let $\boldsymbol{\xi} \triangleq [\xi_1, ..., \xi_m]'$, let $\mathcal{K}^{(i)}$ be an m-by-m matrix with elements

$$\left\{\mathcal{K}^{(i)}\right\}_{s,\tilde{s}} \triangleq \left(\alpha_{i,s} + \beta_{i,s}\right) \frac{\pi_{\tilde{s}}}{\pi_s} \left\{A^i\right\}_{\tilde{s},s} , \quad s, \tilde{s} = 1, ..., m , \tag{3.12}$$

and let $\mathbf{h}_t \triangleq \left[E_{t-1}\left(\sigma_{t,1}^2 \mid S_t = 1\right), ..., E_{t-1}\left(\sigma_{t,m}^2 \mid S_t = m\right)\right]'$ be an m-by-1 vector of the expected state dependent conditional variances. Then, we have

$$\mathbf{h}_t = \boldsymbol{\xi} + \sum_{i=1}^{r} \mathcal{K}^{(i)} \mathbf{h}_{t-i} . \tag{3.13}$$

Define the rm-by-1 vectors $\tilde{\mathbf{h}}_t \triangleq \left[\mathbf{h}_t',\ \mathbf{h}_{t-1}', ...,\ \mathbf{h}_{t-r+1}'\right]'$ and $\tilde{\boldsymbol{\xi}} \triangleq [\boldsymbol{\xi}', 0, ..., 0]'$, and let

$$\Psi_I \triangleq \begin{bmatrix} \mathcal{K}^{(1)} & \mathcal{K}^{(2)} & & \cdots & & \mathcal{K}^{(r)} \\ I_m & 0_m & & \cdots & & 0_m \\ 0_m & I_m & & & & \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ & & & & & \\ 0_m & \cdots & & 0 & I_m & 0_m \end{bmatrix} \tag{3.14}$$

be an mr-by-mr matrix where I_m represents the identity matrix of size m-by-m and 0_m is an m-by-m matrix of zeros. Then a recursive vector form of the expected conditional variance (3.13) can be written as

$$\tilde{\mathbf{h}}_t = \tilde{\boldsymbol{\xi}} + \Psi_I \tilde{\mathbf{h}}_{t-1}, \quad t \geq 0 , \tag{3.15}$$

with some initial conditions $\tilde{\mathbf{h}}_{-1}$.

Let $\rho\left(\cdot\right)$ denote the spectral radius of a matrix, *i.e.,* its largest eigenvalue in modulus, and let Λ_I be an m-by-m square matrix built from the mr-by-mr matrix $\left(I - \Psi_I\right)^{-1}$ such that $\left\{\Lambda_I\right\}_{ij} = \left\{\left(I - \Psi_I\right)^{-1}\right\}_{ij}$, $i, j = 1, ..., m$. Then we have the following theorem:

Theorem 3.1. *An MSG-I process as defined by (3.1) and (3.4) is asymptotically wide-sense stationary with variance* $\lim_{t\to\infty} E\left(\varepsilon_t^2\right) = \boldsymbol{\pi}'\Lambda_I\boldsymbol{\xi}$*, if and only if* $\rho\left(\Psi_I\right) < 1$[2].

Proof. The recursive equation (3.15) can be written as

$$\tilde{\mathbf{h}}_t = \Psi_I^t\tilde{\mathbf{h}}_0 + \sum_{i=0}^{t-1}\Psi_I^t\tilde{\boldsymbol{\xi}}\,, \qquad t \geq 0\,. \tag{3.16}$$

According to the matrix convergence theorem (*e.g.,* [124, pp. 327-329]), a necessary and sufficient condition for the convergence of (3.16) for $t \to \infty$ is $\rho\left(\Psi_I\right) < 1$. Under this condition, Ψ_I^t converges to zero as t goes to infinity and $\sum_{i=0}^{t-1}\Psi_I^i$ converges to $\left(I - \Psi_I\right)^{-1}$, where the matrix $\left(I - \Psi_I\right)$ is then guaranteed to be invertible. Therefore, if $\rho\left(\Psi_I\right) < 1$, equation (3.16) yields

$$\lim_{t\to\infty}\tilde{\mathbf{h}}_t = \left(I - \Psi_I\right)^{-1}\tilde{\boldsymbol{\xi}}\,. \tag{3.17}$$

By definition, the first m elements of $\tilde{\mathbf{h}}_t$ constitute the vector $\mathbf{h}_t$, while the first m elements of $\tilde{\boldsymbol{\xi}}$ constitute the vector $\boldsymbol{\xi}$, and the remaining elements of $\tilde{\boldsymbol{\xi}}$ are zeros. Consequently,

$$\lim_{t\to\infty}\mathbf{h}_t = \Lambda_I\boldsymbol{\xi} \tag{3.18}$$

and using (3.5) we have

$$\lim_{t\to\infty}E\left(\varepsilon_t^2\right) = \boldsymbol{\pi}'\Lambda_I\boldsymbol{\xi}\,. \tag{3.19}$$

Otherwise, if $\rho\left(\Psi_I\right) \geq 1$, the expected variance goes to infinity with the growth of the time index. □

3.2.2 MSG-II model

Another variant of Markov-switching GARCH model has recently been proposed by Haas, Mittnik and Paolella [8]. This model assumes that a Markov chain controls the ARCH parameters at each regime (*i.e.*, ξ_s and $\alpha_{i,s}$), while the *autoregressive* behavior in each regime is subject to the assumption that past conditional variances are in the same regime as that of the current conditional variance. Specifically, the vector of conditional variances $\boldsymbol{\sigma}_t^2 \triangleq \left[\sigma_{t,1}^2, \sigma_{t,2}^2, ..., \sigma_{t,m}^2\right]'$ is given by

[2]Note that for a matrix with nonnegative elements, there exists a real eigenvalue which is equal to the spectral radius [123, p. 288].

$$\boldsymbol{\sigma}_t^2 = \boldsymbol{\xi} + \sum_{i=1}^{q} \boldsymbol{\alpha}_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} B^{(j)} \boldsymbol{\sigma}_{t-j}^2 \ , \tag{3.20}$$

where $\boldsymbol{\alpha}_i \triangleq [\alpha_{i,1}, ..., \alpha_{i,m}]'$, $i = 1, ..., q$, and $\boldsymbol{\beta}_j \triangleq [\beta_{j,1}, ..., \beta_{j,m}]'$, $j = 1, ..., p$, are vectors of state dependent GARCH parameters, and $B^{(j)} \triangleq \text{diag}\{\boldsymbol{\beta}_j\}$ is a diagonal matrix with elements β_j on its diagonal. The same constraints which are sufficient to ensure a positive conditional variance in MSG-I model (3.3) are also applied here to guaranty the positivity of the conditional variance.

Note that the conditional variance at a specific regime depends on previous conditional variances of the same regime through the diagonal matrices $B^{(j)}$. Consequently, this model allows derivation of the conditional variance at a given time from past observations only. Furthermore, the MSG-II model is analytically more tractable than MSG-I model [8] and its conditional variance can be straightforwardly constructed since the conditional variance at a specific time does not depend on previous state probabilities but only on previous observations and previous conditional variances.

Let $\boldsymbol{\alpha}_i$ be an m-by-1 vector of zeros for $i > q$ and let $B^{(j)} = 0_m$ for $j > p$. Let $\Omega^{(i)}$ denote an m^2-by-m^2 block matrix of basic dimension m-by-m

$$\Omega^{(i)} \triangleq \begin{bmatrix} \Omega_{11}^{(i)} & \Omega_{21}^{(i)} & \cdots & \Omega_{m1}^{(i)} \\ \Omega_{12}^{(i)} & \Omega_{22}^{(i)} & \cdots & \Omega_{m2}^{(i)} \\ \vdots & & & \vdots \\ \Omega_{1m}^{(i)} & \Omega_{2m}^{(i)} & \cdots & \Omega_{mm}^{(i)} \end{bmatrix} , \tag{3.21}$$

with each block given by

$$\Omega_{s\tilde{s}}^{(i)} \triangleq p\left(S_{t-i} = s \,|\, S_t = \tilde{s}\right) \left(\boldsymbol{\alpha}_i \mathbf{e}_s' + \mathbf{B}^{(i)}\right), \quad s, \tilde{s} = 1, ..., m \,, \tag{3.22}$$

where $\mathbf{e}_s$ is an m-by-1 vector of all zeros, except its sth element which is one. We define an rm^2-by-rm^2 matrix by

$$\Psi_{II} \triangleq \begin{bmatrix} \Omega^{(1)} & \Omega^{(2)} & & \cdots & & \Omega^{(r)} \\ I_{m^2} & 0_{m^2} & & \cdots & & 0_{m^2} \\ 0_{m^2} & I_{m^2} & & & & \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ & & & & & \\ 0_{m^2} & \cdots & & 0_{m^2} & I_{m^2} & 0_{m^2} \end{bmatrix}. \tag{3.23}$$

Let Λ_{II} be an m^2-by-m^2 matrix which is built from the rm^2-by-rm^2 matrix $(I - \Psi_{II})^{-1}$ such that $\{\Lambda_{II}\}_{ij} = \left\{(I - \Psi_{II})^{-1}\right\}_{ij}$, $i, j = 1, ..., m^2$. Let $\bar{\boldsymbol{\pi}} \triangleq [\pi_1 \mathbf{e}_1', \pi_2 \mathbf{e}_2', ..., \pi_m \mathbf{e}_m']'$, then we get the following theorem for the stationarity condition of an MSG-II process:

Theorem 3.2. *An MSG-II process as defined by (3.1) and (3.20) is asymptotically wide-sense stationary with variance* $\lim_{t\to\infty} E(\varepsilon_t^2) = \bar{\boldsymbol{\pi}}' \Lambda_{II} \boldsymbol{\xi}$*, if and only if* $\rho(\Psi_{II}) < 1$.

The proof is given in Appendix 3.A.

3.2.3 Comparison of stationarity conditions

It has been pointed out in [8], that stationarity of the MSG-II model with $p = q = 1$ requires that the regression parameters $\beta_{1,s} < 1$ for all s. It follows from (3.20) that for general order (p, q), it is necessary that $\sum_{i=1}^{m} \beta_{i,s} < 1$. However, the reaction parameters $\alpha_{i,s}$ may become rather large with correspondence to the regime probabilities. For MSG-I and model, the reaction parameters $\alpha_{i,s}$ as well as the regression parameters $\beta_{i,s}$ may be larger than one, provided that the corresponding regime probabilities are sufficiently small. Furthermore, in the representative matrix Ψ_I (3.14) the reaction parameters and the regression parameters are weighted by the same weights $p(S_{t-i} = s \mid S_t = \tilde{s})$. Consequently, for a given state s, the values of $\alpha_{i,s}$ and $\beta_{i,s}$ in MSG-I model have the same contribution to the model stationarity[3], but for MSG-II model, each of them affects differently the heteroscedasticity evolution. Figure 3.1 illustrates the stationarity regions for MSG-I model (solid line) and MSG-II (dashed-dotted line), in the case of two-state Markov chains and GARCH of order $(1, 1)$. In (a), the regime transition probabilities are

[3]This also holds for the natural extension of GARCH(p, q) to Markov-switching, which has been analyzed in [122].

$a_{1,1} = 0.6$ and $a_{2,2} = 0.7$ and the reaction parameters are $\alpha_{1,1} = 0.4$ and $\alpha_{1,2} = 0.5$. The stationarity region is the interior intersection of each curve and the two axes. In (b), $a_{1,1} = 0.2$, $a_{2,2} = 0.3$ are considered with reaction parameters $\alpha_{1,1} = 0.8$ and $\alpha_{1,2} = 0.2$. For the MSG-I model, stationarity is allowed with regression parameters larger than one while for the MSG-II, $\beta_{1,1}$ and $\beta_{1,2}$ must be both smaller than one for stationarity. In both cases, $\pi_2 > \pi_1$, however, in (a) the stationarity region of the MSG-II is contained in the stationarity region of MSG-I while in (b), in which case $\alpha_{1,1} >> \alpha_{1,2}$, for $\beta_{1,1} \in [0.2, 0.55]$ stationarity is achieved with a larger $\beta_{1,2}$ for the MSG-II than for the MSG-I.

3.3 Relation to other works

Klaassen [7] developed conditions which are necessary, but not necessarily sufficient, for asymptotic stationarity of a two-state MSG-I model of order $(1, 1)$. Consider the 2-by-2 matrix C with elements

$$c_{ij} = a_{ji} \left(\alpha_{1,i} + \beta_{1,i}\right) \pi_j / \pi_i, \quad i, j = 1, 2\,, \tag{3.24}$$

Klaassen showed that the stationary variance of the process is given by

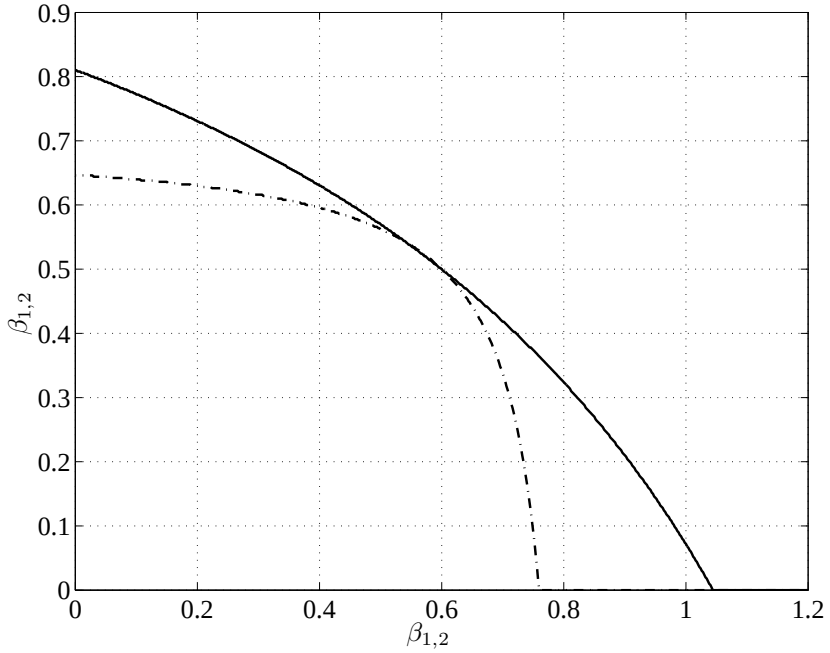
$$\sigma^2 = \boldsymbol{\pi}'(I - C)^{-1}\boldsymbol{\xi}\,, \tag{3.25}$$

and that the conditions:

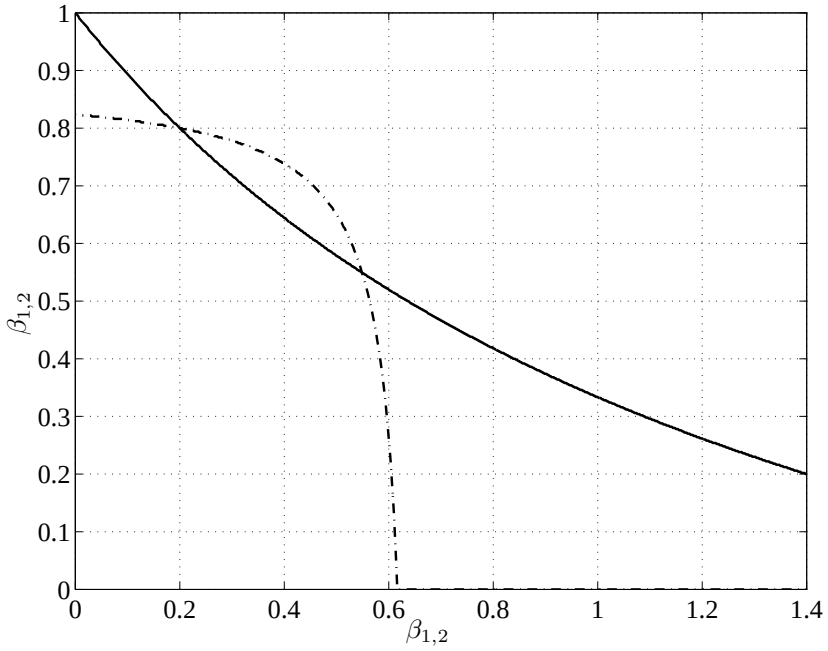
$$c_{11}, c_{22} < 1, \text{ and } \det(I - C) > 0\,, \tag{3.26}$$

are necessary to ensure that the stationary variance is finite and positive.

For the special case of our analysis for GARCH orders of $(1, 1)$ and MSG-I model with two states, the representative matrix Ψ_I reduces to matrix C and the stationary variance reduces to the expression given in (3.25). Metzler showed [125] that for a nonnegative matrix C (*i.e.*, $c_{ij} \geq 0$), $\rho\,(C) < 1$ if and only if all of the principal minors of $(I - C)$ are positive. Furthermore, together with Hawkins-Simon condition [126], $\rho\,(C)$ is less than one if and only if $(I - C)^{-1}$ has no negative elements. Therefore, for the nonnegative matrix C, the condition $c_{11}, c_{22} < 1$ implies $\det(I - C) > 0$ and it is equivalent to $\rho\,(C) < 1$. Accordingly, the conditions of Klaassen are not only necessary but also sufficient for asymptotic stationarity.



(a)



(b)

Figure 3.1: Stationarity regions for two-state Markov-chains with GARCH of order $(1, 1)$ corresponding to MSG-I (solid line) and MSG-II (dashed-dotted line). The regime transition probabilities and the reaction parameters are (a) $a_{1,1} = 0.6$, $a_{2,2} = 0.7$ and $\alpha_{1,1} = 0.4$, $\alpha_{1,2} = 0.5$; (b) $a_{1,1} = 0.2$, $a_{2,2} = 0.3$ and $\alpha_{1,1} = 0.8$, $\alpha_{1,2} = 0.2$.

A necessary and sufficient condition for asymptotic stationarity of an MSG-II model of order $(1,1)$ has been developed by Haas *et al.* [8]. Accordingly, the largest eigenvalue in modulus of an m^2-by-m^2 block matrix D is constrained to be less than one, where

$$D = \begin{bmatrix} D_{11} & D_{21} & \cdots & D_{m1} \\ D_{12} & D_{22} & \cdots & D_{m2} \\ \vdots & \vdots & & \vdots \\ D_{1m} & D_{2m} & \cdots & D_{mm} \end{bmatrix} \tag{3.27}$$

is built from matrices D_{ij} of size m-by-m which are obtained by

$$D_{ij} = a_{ij} \left(B^{(1)} + \boldsymbol{\alpha}_1 \mathbf{e}_j' \right) . \tag{3.28}$$

The stationarity analysis in [8] for an MSG-II process employs a forward recursive calculation of the expected conditional variance, assuming some initial conditions. As a result, the probabilities of state transitions, a_{ij}, are used for evaluating the expectation of the one step ahead conditional variance. Our analysis manipulates a backward recursion of the conditional variance expectation, and thus, it uses the stationary probabilities of the Markov chain, along with the transition probabilities, to generate previous conditional states probabilities $p\left(s_{t-i} \,|\, s_t\right)$. Therefore, when we degenerate the MSG-II model to order $p = q = 1$ (which is the case analyzed in [8]) the block matrices D (3.27) and Ψ_{II} (3.23) are not identical, and specifically, for that order of model we have $\Psi_{II} = \Omega^{(1)}$ and

$$\Omega_{ij}^{(1)} = \frac{\pi_i a_{ij}}{\pi_j a_{ji}} D_{ji} \, . \tag{3.29}$$

Although our representative matrix Ψ_{II} and that developed in [8] do not share the same elements, we show in Appendix 3.B that their eigenvalues are identical and therefore both conditions are equivalent for that order of MSG-II.

A special case of any of the MSG models is a degenerated case of having a single regime of order (p, q) (the models reduce to a standard GARCH(p, q) model). In that case, the representative matrices are equal, $\Psi_I = \Psi_{II}$. Francq, Roussignol and Zakoïan [122] developed a stationarity condition for the *natural* case of Markov-switching GARCH model, in which case the conditional variance depends on the active regime-path. For the special case of a single-regime model they got the transition matrix of a standard GARCH(p, q) model which is equal to that which is derived *e.g.*, by substituting $\mathcal{K}^{(i)} =$

$\alpha_{i,1}+\beta_{i,1}$ and $I_1 = 1$ in (3.14). They showed that having the spectral radius of that matrix to be less than one is equivalent to Bollerslev's condition for the asymptotic wide-sense stationarity of a GARCH(p, q) model, $\sum_{i=1}^{r} (\alpha_{i,1} + \beta_{i,1}) < 1$ [1].

3.4 Conclusions

Conditions for asymptotic wide-sense stationarity of random processes with time-variant distributions are useful for ensuring the existence of a finite asymptotic volatility of the process. We developed a comprehensive approach for stationarity analysis of Markov-switching GARCH processes where finite state space Markov chains control the switching between regimes, and GARCH models of order (p, q) are active in each regime. Necessary and sufficient conditions for the asymptotic stationarity are obtained by constraining the spectral radius of representative matrices, which are built from the model parameters. These matrices also enable derivation of compact expressions for the stationary variance of the processes.

3.A Proof of Theorem 3.2

In this appendix we prove Theorem 3.2, which gives necessary and sufficient condition for the asymptotic wide-sense stationarity of MSG-II model and also its stationary variance.

Following (3.20) and (3.5), the expectation of the MSG-II conditional variance under a chain state s, follows:

$$E_{t-1}(\sigma_{t,s}^2 \,|\, s_t) = \xi_s + \sum_{i=1}^{q} \alpha_{i,s} E_{t-1}\left(\varepsilon_{t-i}^2 \,|\, s_t\right) + \sum_{j=1}^{p} \beta_{j,s} E_{t-1}(\sigma_{t-j,s}^2 \,|\, s_t) \tag{3.30}$$

where using (3.7) and (3.8)

$$E_{t-1}\left(\varepsilon_{t-i}^2 \,|\, s_t\right) = \sum_{s_{t-i=1}}^{m} p\left(s_{t-i} \,|\, s_t\right) E_{t-i-1}\left(\sigma_{t-i,s_{t-i}}^2 \,|\, s_{t-i}\right), \tag{3.31}$$

and

$$\begin{aligned} E_{t-1}(\sigma_{t-j,s}^2 \,|\, s_t) &= E_{S_{t-j}}\left[E_{t-1}(\sigma_{t-j,s}^2 \,|\, s_{t-j}, s_t)\right] \\ &= \sum_{s_{t-j=1}}^{m} p\left(s_{t-j} \,|\, s_t\right) E_{t-i-1}\left(\sigma_{t-j,s}^2 \,|\, s_{t-j}\right). \end{aligned} \tag{3.32}$$

The main difference between an MSG-II model and MSG-I model is that the conditional variance depends on previous conditional variances of the same regime, regardless of the past regimes path. By contrast, for MSG-I model, the conditional variance is a linear combination of past state-dependent conditional variances, where for each one the state is conditioned to be the active one. Consequently, the computation of the unconditional variance for an MSG-II model requires the terms $E_{t-j-1}(\sigma^2_{t-j,s} \,|\, s_{t-j})$ for all $s = 1, ..., m$, while in case of MSG-I model, only $E_{t-j-1}(\sigma^2_{t-j,s_{t-j}} \,|\, s_{t-j})$ is relevant to calculate the expectation of the unconditional variance. Accordingly, an m^2-by-1 vector is necessary to represent $E_{t-1}(\sigma^2_{t,s} \,|\, s_t)$ elements, and rm^2-by-rm^2 matrix is employed for the recursive formulation.

By substituting (3.32) and (3.31) into (3.30) we have

$$E_{t-1}(\sigma^2_{t,s} \,|\, s_t) = \xi_s + \sum_{i=1}^{r} \sum_{s_{t-i}=1}^{m} p\left(s_{t-i} \,|\, s_t\right) \left[\alpha_{i,s} E_{t-i-1}\left(\sigma^2_{t-i,s_{t-i}} \,|\, s_{t-i}\right) + \beta_{i,s} E_{t-i-1}\left(\sigma^2_{t-i,s} \,|\, s_{t-i}\right)\right] . \tag{3.33}$$

Let $g_t\left(s, s_t\right) \triangleq E_{t-1}\left(\sigma^2_{t,s} \,|\, s_t\right)$, and let $\mathbf{g}_t \triangleq \left[g_t\left(1,1\right), g_t\left(2,1\right), ..., g_t\left(m,1\right), g_t\left(1,2\right), ..., g_t\left(m,m\right)\right]'$ be a vector of expected, state dependent, conditional variances. Then, a recursive formulation of the conditional variance is given by

$$\tilde{\mathbf{g}}_t = \tilde{\boldsymbol{\xi}} + \Psi_{II} \tilde{\mathbf{g}}_{t-1}, \quad t \geq 0\,, \tag{3.34}$$

where $\tilde{\mathbf{g}}_t \triangleq \left[\mathbf{g}'_t, \mathbf{g}'_{t-1}, ..., \mathbf{g}'_{t-r+1}\right]'$. The completion of this proof follows the proof of Theorem 1.

3.B Equivalence with Haas condition

In this appendix we show that the eigenvalues of matrices D (3.27) and $\Psi_{II} = \Omega^{(1)}$ (3.23) are equal for the case of an m-state MSG-II model of order $(1,1)$.

Let $\tilde{D}$ denote an m^2-by-m^2 matrix that is given by

$$\tilde{D} \triangleq \begin{bmatrix} B^{(1)} + \boldsymbol{\alpha}_1 \mathbf{e}'_1 & 0_m & \cdots & 0_m \\ 0_m & B^{(1)} + \boldsymbol{\alpha}_1 \mathbf{e}'_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0_m \\ 0_m & \cdots & 0_m & B^{(1)} + \boldsymbol{\alpha}_1 \mathbf{e}'_m \end{bmatrix}, \tag{3.35}$$

and let $\otimes$ denote the Kronecker product. Then $D = \tilde{D}\,(A' \otimes I_m)$. Let $\mathbf{1}_m$ denote an m-by-1 vector of ones and let $P \triangleq \mathrm{diag}(\boldsymbol{\pi} \otimes \mathbf{1}_m)$. By substituting (3.28) into (3.29), we have

$$\Omega_{ij}^{(1)} = \frac{\pi_i}{\pi_j} a_{ij} \left(B^{(1)} + \boldsymbol{\alpha}_1 \mathbf{e}_i' \right) \tag{3.36}$$

and

$$\Omega^{(1)} = P^{-1} \left(A' \otimes I_m \right) \tilde{D} P. \tag{3.37}$$

Therefore, $\Omega^{(1)}$ and $(A' \otimes I_m)\,\tilde{D}$ are similar matrices, and the spectrum of $\Omega^{(1)}$, $\mathrm{eig}\{\Omega^{(1)}\}$, satisfies

$$\mathrm{eig}\left\{\Omega^{(1)}\right\} = \mathrm{eig}\left\{ \left(A' \otimes I_m \right) \tilde{D} \right\} = \mathrm{eig}\left\{ \tilde{D} \left(A' \otimes I_m \right) \right\} = \mathrm{eig}\left\{ D \right\} . \tag{3.38}$$

Chapter 4

Markov-Switching GARCH Process in the Short-Time Fourier Transform Domain[1]

In this chapter, we introduce a Markov-switching generalized autoregressive conditional heteroscedasticity (GARCH) model for nonstationary processes with time-varying volatility structure in the short-time Fourier transform (STFT) domain. The expansion coefficients in the STFT domain are modeled as a multivariate complex GARCH process with Markov-switching regimes. The GARCH formulation parameterizes the correlation between sequential conditional variances while the Markov chain allows the process to switch between regimes of different GARCH formulations. We obtain a necessary and sufficient condition for the asymptotic wide-sense stationarity of the model, and develop a recursive algorithm for signal restoration in a noisy environment. The conditional variance is estimated by iterating propagation and update steps with regime conditional probabilities, while the model parameters are evaluated a priori from a training data set. Experimental results demonstrate the performance of the proposed algorithm.

In Appendix 4.A, we introduce an application of the Markov-switching GARCH model in the STFT domain to speech enhancement. A GARCH model is utilized with Markov switching regimes, where the parameters are assumed to be frequency variant. The model parameters are evaluated in each frequency subband and a special state (regime) is de-

[1]This chapter is based on [127, 128].

fined for the case where speech coefficients are absent or below a threshold level. The problem of speech enhancement under speech presence uncertainty is addressed and it is shown a soft voice activity detector may be inherently incorporated within the algorithm. Experimental results demonstrate the potential of our proposed model to improve noise reduction while retaining weak components of the speech signal.

4.1 Introduction

The generalized autoregressive conditional heteroscedasticity (GARCH) model is widely-used in the field of econometrics for volatility forecast derivation of economic rates. This model, first introduced by Bollerslev [1] as a generalization of the ARCH model [2], explicitly parameterizes the time-varying volatility by using both recent conditional variances and recent squared innovations. GARCH models preserve the persistence of the process volatility in the sense that small variations tend to follow small variations and large variations tend to follow large variations. Incorporating GARCH models with hidden Markov chains, where each state (regime) of the chain implies a different GARCH behavior, extends the dynamic formulation of the model and enables a better fit for a process with a more complex time-varying volatility structure [7–9]. However, a major drawback of such models is that estimating the volatility with switching-regimes requires knowledge of the entire history of the process, including the regime path. Consequently, Cai [12] and Hamilton and Susmel [13] proposed a Markov-switching ARCH model, which avoids problems of path dependency in a noiseless environment. The conditional variance in ARCH models depends on previous observations only, so the Markov chain does not have to be known for constructing the conditional variance for a given regime. Gray [6] introduced a variant of Markov-switching GARCH model relying on the assumption that the conditional variance given current regime is dependent on the *expectation* of the previous conditional variances rather than their values. Accordingly, the conditional variance depends on some finite, state dependent, expected conditional variances via their conditional state probabilities. Klaassen [7] proposed modifying Gray's model by manipulating the current regime and all available observations while evaluating the expectation of previous conditional variances. A different method for reducing the dependency of the conditional variance on

past regimes has recently been proposed by Haas, Mittnik and Paolella [8]. Accordingly, a Markov chain governs the ARCH parameters while the autoregressive behavior of the conditional variance is subject to the assumption that past conditional variances are in the same regime as that of the current conditional variance. Gray, Klaassen and Haas *et al.* developed their variants of Markov-switching GARCH models for improved volatility forecasts of financial time-series under possible existence of shocks. They assumed that a process is observed in a noiseless environment so that its past observations provide a complete specification of its current conditional variance, for any given regime.

Recently, GARCH models have been employed for modeling speech signals in the time-frequency domain [23–25]. Speech signals in the short-time Fourier transform (STFT) domain demonstrate both "variability clustering" and heavy tail behavior similarly to financial time-series [25]. Motivated by these characteristics, it was proposed to model the conditional variance of speech signals in the STFT domain by a complex, K-dimensional GARCH model, with statistically independent elements (given past information) sharing the same GARCH specification. This time-frequency GARCH (TF-GARCH) model has been shown useful for speech enhancement applications, but it relies on the assumption that the model parameters are time-invariant. In [26], a GARCH model has been utilized in the time domain for speech recognition applications. The model parameters, characterizing the speech phonemes, are assumed speaker independent and time-varying. It was shown that estimating the GARCH specifications for each speech segment and using the parameters as part of the signal characteristics, speech recognition performance can be improved.

In this chapter, we introduce a Markov-switching time-frequency GARCH (MSTF-GARCH) model which exploits the advantages of both the conditional heteroscedasticity structure of GARCH models and the time-varying characteristics of hidden Markov chains. Modeling probability density functions of speech signals by utilizing hidden Markov models has been found useful in speech recognition applications [63–65], and modeling the speech spectral coefficients as hidden Markov processes with a probability density prototype in each frame was applied to the problem of speech enhancement [35, 62]. Here we model the expansion coefficients of nonstationary random signals in the time-frequency domain as multivariate complex GARCH processes with Markov-switching regimes, and

obtain a necessary and sufficient condition for the asymptotic wide-sense stationarity of the model. A corresponding recursive algorithm is developed for signal restoration in a noisy environment. The conditional variance is estimated by iterating propagation and update steps with regime conditional probabilities. The model parameters are estimated from a training data set prior to the signal restoration using maximum-likelihood (ML) approach, and the number of states is assumed to be known. We show that the derivation in [117] of bounds on the mean-square error (MSE) of a composite source signal estimation is applicable for obtaining an upper bound on the MSE of a single step MSTF-GARCH estimation. Experimental results demonstrate the improved performance of the proposed algorithm for restoration of MSTF-GARCH process compared to using an estimator which assumes a stationary process and compared to using an estimator which assumes a smaller number of regimes than the process actually has. Furthermore, it is demonstrated that the squared absolute values of speech coefficients in the STFT domain are better evaluated by using the MSTF-GARCH model than by using the decision-directed approach.

This chapter is organized as follows. In Section 4.2, we introduce the Markov-switching time-frequency GARCH model and obtain a necessary and sufficient condition for its asymptotic wide-sense stationarity. In Section 4.3, we address the problem of signal estimation from noisy observations. In Section 4.4, we derive an upper bound on a single estimation step mean-square error. In Section 4.5, we address the problem of model estimation. Finally, in Section 4.6 we provide some experimental results which demonstrate restoration of MSTF-GARCH process from noisy observations, and estimation of conditional variances and squared absolute values in the STFT domain from noisy speech signals.

4.2 Markov-switching time-frequency GARCH model

In this section we briefly review the TF-GARCH model [25], and introduce a new time-frequency GARCH model with Markov-switching regimes, which allows further flexibility in the formulation of the time variation of the conditional variance.

4.2.1 Time-frequency GARCH model

Let $\{X_{tk} \,|\, t = 0, ..., T-1\,,\, k = 0, ..., K-1\}$ be the coefficients of a time-frequency transformation of a discrete-time signal x (*e.g.,* STFT coefficients), where t is the time frame index and k is the frequency-bin index. Let $\mathbf{X}_t \triangleq [X_{t,0}, ..., X_{t,K-1}]'$ be the vector of spectral coefficients at time frame t, let $\mathcal{X}^\tau = \mathcal{X}_0^\tau \triangleq \{\mathbf{X}_t \,|\, t = 0, ..., \tau\}$ represent the set of spectral coefficients up to time τ, and let $\lambda_{tk|\tau} \triangleq E\{|X_{tk}|^2 \,|\, \mathcal{X}^\tau\}$ denote the *conditional variance* of the spectral coefficient at time-frequency bin (t, k), given the clean spectral coefficients up to time τ. Let $\{\mathbf{V}_t\} \in \mathbb{C}^K$ be a complex Gaussian random process with $\mathbf{V}_t \sim \mathcal{CN}(0, I_K)$, where I_K is a K-by-K identity matrix. A K-dimensional time-frequency GARCH model of order (p, q), is defined as follows [25]:

$$X_{tk} = \sqrt{\lambda_{tk|t-1}} V_{tk}\,, \qquad k = 0, ..., K-1 \tag{4.1}$$

$$\boldsymbol{\lambda}_{t|t-1} = \zeta \cdot \mathbf{1} + \sum_{i=1}^{q} \alpha_i \mathbf{X}_{t-i} \odot \mathbf{X}_{t-i}^* + \sum_{j=1}^{p} \beta_j \boldsymbol{\lambda}_{t-j|t-j-1}\,, \tag{4.2}$$

where $\mathbf{1}$ denotes a vector of ones, $\odot$ denotes a term-by-term multiplication and * denotes complex conjugation. The conditional variance vector, $\boldsymbol{\lambda}_{t|t-1} = E\{\mathbf{X}_t \odot \mathbf{X}_t^* \,|\, \mathcal{X}^{t-1}\}$, referred to as the *one-frame-ahead conditional variance* [25], is a linear function of the coefficients' past squared values and conditional variances, where

$$\zeta > 0, \qquad \alpha_i \geq 0, \quad i = 1, ..., q, \qquad \beta_j \geq 0, \quad j = 1, ..., p\,, \tag{4.3}$$

are sufficient constraints for the positivity of the conditional variance [1]. The time-frequency GARCH has been introduced in [23] for modeling speech signals in the STFT domain, but the parameters of the GARCH model are assumed time invariant. Extending this model such that the model parameters may vary with time introduces additional flexibility in the model formulation, which may result in better characterization of speech signals and improved restoration in noisy environments.

4.2.2 MSTF-GARCH formulation

Let S_t denote the (unobserved) state at time t and let s_t be a realization of S_t, assuming S_t is a first-order Markov chain. Let $\mathcal{I}^t \triangleq \{\mathcal{X}^t, \mathcal{S}^t\}$ denote all available information up to time t, which contains the clean signal coefficients and the regimes path up to

time t, $\mathcal{S}^t \triangleq \{s_0, ..., s_t\}$. Denote by $\lambda_{tk|t-1,s_t} \triangleq E\left\{|X_{tk}|^2 \,|\, \mathcal{I}^{t-1}, s_t\right\}$ the one-frame-ahead conditional variance of the spectral coefficient X_{tk} given the information up to time $t-1$ and the chain state s_t. We assume that the spectral coefficients X_{tk} are generated by an m-state Markov-switching time-frequency GARCH process of order (p, q), denoted by $X_{tk} \sim$MSTF-GARCH(p, q), which follows:

$$X_{tk} = \sqrt{\lambda_{tk|t-1,s_t}} V_{tk}\,, \quad k = 0, ..., K-1\,, \tag{4.4}$$

and the one-frame-ahead conditional variance evolves as follows:

$$\boldsymbol{\lambda}_{t|t-1,s_t} = \zeta_{s_t}\mathbf{1} + \sum_{i=1}^{q} \alpha_{i,s_t} \mathbf{X}_{t-i} \odot \mathbf{X}^*_{t-i} + \sum_{j=1}^{p} \beta_{j,s_t} \boldsymbol{\lambda}_{t-j|t-j-1,s_{t-j}}\,, \tag{4.5}$$

where

$$\zeta_s > 0, \quad \alpha_{i,s} \geq 0, \quad \beta_{j,s} \geq 0, \qquad i = 1, ..., q, \qquad j = 1, ..., p, \qquad s = 1, ..., m \tag{4.6}$$

are sufficient constraints for the positivity of the one-frame-ahead conditional variance. It follows from (4.4) and (4.5) that the conditional density of the coefficients depends on past values (through previous conditional variances) and also on the regime-path up to the current time. As considered in previous works on TF-GARCH, we assume that the model parameters are frequency-invariant. This restriction can be easily relaxed for the case of frequency (or sub-band) dependent parameters, *i.e.*, $\zeta_{k,s}$, $\alpha_{i,k,s}$ and $\beta_{i,k,s}$, but the complexity of the model estimation then grows rapidly (see Section 4.5).

GARCH models provide a rich class of possible parametrization of conditional heteroscedasticity (*i.e.*, time-varying volatility) and the hidden Markov chain allows these GARCH formulations to switch along time. Volatility persistence naturally arises in a single-regime GARCH model. However, the existence of a Markov chain with different GARCH parameters allows the process to switch between regimes of different volatility formulations and different levels of volatility.

4.2.3 Stationarity of an MSTF-GARCH process

The conditional variance of a GARCH process, and in particular of a Markov-switching GARCH process, changes recursively over time. Consequently, asymptotic wide-sence

stationarity is required to ensure a finite second-order moment [7, 8, 121]. Necessary and sufficient conditions for the asymptotic stationarity of three variants of GARCH models with Markov-switching regimes have been derived in [121]. Those models generalize the models of Gray [6], Klaassen [7] and Haas *et al.* [8], but they all differ from our MSTF-GARCH model, which is a multivariate, complex valued process that entails the regime path for the construction of the conditional variance from past observations. A necessary and sufficient condition for asymptotic wide-sense stationarity of an MSTF-GARCH process has been derived in [128]. For the completeness of this chapter we briefly summarize these results:

Assuming a stationary Markov chain with stationary probabilities $\pi_s = p\left(S_t = s\right)$, the unconditional variance of the process can be calculated using (4.4) and (4.5):

$$\begin{aligned} E\left\{\mathbf{X}_t \odot \mathbf{X}_t^*\right\} &= \sum_{s_t} \pi_{s_t} E\left\{\mathbf{X}_t \odot \mathbf{X}_t^* \,|\, s_t\right\} \\ &= \sum_{s_t} \pi_{s_t} E\left\{\boldsymbol{\lambda}_{t|t-1,s_t}\right\} , \end{aligned} \tag{4.7}$$

where

$$E\left\{\boldsymbol{\lambda}_{t|t-1,s_t}\right\} = \zeta_{s_t}\mathbf{1} + \sum_{i=1}^{q} \alpha_{i,s_t} E\left\{\mathbf{X}_{t-i} \odot \mathbf{X}_{t-i}^* \,|\, s_t\right\} + \sum_{j=1}^{p} \beta_{j,s_t} E\left\{\boldsymbol{\lambda}_{t-j|t-j-1,S_{t-j}} \,|\, s_t\right\} , \tag{4.8}$$

and

$$E\left\{\boldsymbol{\lambda}_{t-i|t-i-1,S_{t-i}} \,|\, s_t\right\} = \sum_{s_{t-i}} p\left(s_{t-i} \,|\, s_t\right) E\left\{\boldsymbol{\lambda}_{t-i|t-i-1,s_{t-i}}\right\} . \tag{4.9}$$

Note that $E\left\{\boldsymbol{\lambda}_{t|t-1,s_t}\right\}$ denotes the expected value of the conditional variance under the regime $S_t = s_t$, but $E\left\{\boldsymbol{\lambda}_{t|t-1,S_t} \,|\, \cdot\right\}$ denotes a conditional expectation of the conditional variance at time t where the active regime at that time is unknown. Since no prior information is given, we have

$$E\left\{\mathbf{X}_{t-i} \odot \mathbf{X}_{t-i}^* \,|\, s_t\right\} = \sum_{s_{t-i}} p\left(s_{t-i} \,|\, s_t\right) E\left\{\boldsymbol{\lambda}_{t-i|t-i-1,s_{t-i}}\right\} , \tag{4.10}$$

and consequently we obtain [128]

$$E\left\{\boldsymbol{\lambda}_{t|t-1,s_t}\right\} = \zeta_{s_t}\mathbf{1} + \sum_{i=1}^{r} \sum_{s_{t-i}} \left(\alpha_{i,s_t} + \beta_{i,s_t}\right) \frac{\pi_{s_{t-i}}}{\pi_{s_t}} \left\{A^i\right\}_{s_{t-i},s_t} \mathbf{E}\left\{\boldsymbol{\lambda}_{t-i|t-i-1,s_{t-i}}\right\} , \tag{4.11}$$

where $r \triangleq \max\{p,q\}$, $\alpha_{i,s_t} \triangleq 0\,\forall i > q$, $\beta_{i,s_t} \triangleq 0\,\forall i > p$, and A is the transition probabilities matrix, *i.e.*, $\{A\}_{ij} \triangleq a_{ij} = p\left(S_t = j \,|\, S_{t-1} = i\right)$. Define m-by-m matrices $\mathcal{K}_i\,,\ i = 1,...,r$ with elements

$$\{\mathcal{K}_i\}_{s,\tilde{s}} \triangleq (\alpha_{i,s} + \beta_{i,s})\,\frac{\pi_{\tilde{s}}}{\pi_s}\left\{A^i\right\}_{\tilde{s},s}\,,\quad s,\tilde{s} = 1,...,m\,, \tag{4.12}$$

and an mr-by-mr matrix as follows

$$\Psi \triangleq \begin{bmatrix} \mathcal{K}_1 & \mathcal{K}_2 & & \cdots & & \mathcal{K}_r \\ I_m & 0 & & \cdots & & 0 \\ 0 & I_m & & & & \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ & & & & & \\ 0 & \cdots & & 0 & I_m & 0 \end{bmatrix}. \tag{4.13}$$

Let $\rho\,(\cdot)$ denote the spectral radius of a matrix, *i.e.*, its largest eigenvalue in modulus, and let Φ be an m-by-m square matrix built from the mr-by-mr matrix $(I-\Psi)^{-1}$ such that $\{\Phi\}_{ij} = \left\{(I-\Psi)^{-1}\right\}_{ij}\,,\ i,j = 1,...,m$. Then a necessary and sufficient condition for asymptotic wide-sense stationarity of an MSTF-GARCH process is $\rho\,(\Psi) < 1$, and the asymptotic covariance matrix of the process is then a diagonal matrix (see [128] for a detailed proof):

$$\lim_{t\to\infty} E\left\{\mathbf{X}_t\mathbf{X}_t^H\right\} = (\boldsymbol{\pi}\Phi\boldsymbol{\zeta})\,I_K\,, \tag{4.14}$$

where $\boldsymbol{\zeta} \triangleq [\zeta_1,...,\zeta_m]'$, $\boldsymbol{\pi}$ is the row vector of the stationary probabilities of the Markov chain, and $(\cdot)^H$ denotes the Hermitian transpose operation.

This stationarity condition is a necessary and sufficient condition for the existence of a finite second-order moment of the process. It implies that in some regimes (but not in all of them) the conditional variance may grow over time (*i.e.*, $\sum_i \alpha_{i,s} + \sum_j \beta_{j,s} > 1$ for some states s) but still the unconditional variance can be finite [121, 128].

4.3 Restoration of noisy MSTF-GARCH process

In this section we develop a recursive algorithm for the restoration of MSTF-GARCH processes observed in additive stationary noise.

A hidden Markov process is a discrete-time finite-state Markov chain observed through a memoryless invariant channel, where the chain state is assumed to be hidden but the transition probabilities between sequential states are assumed to be known. As a consequence of the memoryless channel, the conditional density of the observed signal at time t (say $\mathbf{X}_t$) given the chain state s_t, depends only on the given state and not on previous observations, *i.e.,* the conditional density of a hidden Markov process (HMP) realizes $p(\mathbf{X}_t \,|\, s_t, \mathbf{X}_{t-1}, \mathbf{X}_{t-2}, ...) = p(\mathbf{X}_t \,|\, s_t)$. Combining GARCH models with hidden Markov chains, where each state is assumed to have a different GARCH formulation, introduces further complexity when trying to forecast or estimate the process, since the conditional variance of the process evolves as a function of previous conditional variances, as implied from (4.5). Consequently, the conditional density depends on the entire history of the process, *i.e.*, past values and active states. To avoid this problem, several variants of GARCH processes with Markov-switching regimes have been proposed, *e.g.*, [6–8]. These models formulate differently the conditional variance at any regime as dependent on past signal observations only. However, these variants of Markov-switching GARCH models have been developed for the purpose of forecasting volatility of financial time-series, assuming that the process is observed in a noiseless environment, and that all past clean signal values are given.

We use an MSTF-GARCH$(1,1)$ model, as defined in (4.4) and (4.5), to model complex, nonstationarity random signals and we develop a recursive signal estimation algorithm for restoring the clean signal and its second-order moment, from noisy observations. The order $(1,1)$ is chosen for computational simplicity since higher (p,q)-orders imply strong dependency of successive conditional variances. Therefore, $p = q = 1$ is generally assumed for the applications of Markov-switching GARCH modeling, *e.g.*, [6–8, 12, 13]. Let $\{X_{tk}\}$ and $\{D_{tk}\}$ denote the spectral coefficients of signal and uncorrelated additive noise signal, respectively, and let $Y_{tk} = X_{tk} + D_{tk}$ represent the observed signal. Let $\mathbf{X}_t$ be a K-dimensional complex-valued stochastic process, which evolves as an m-state first-order MSTF-GARCH, *i.e.,* $\mathbf{X}_t \sim$MSTF-GARCH$(1,1)$, and let $\mathbf{D}_t$ represent a K-dimensional complex Gaussian random noise, $\mathbf{D}_t \sim \mathcal{CN}(0, R^d)$, with known diagonal covariance matrix $R^d = \text{diag}\{\boldsymbol{\sigma}^2\}$. We assume that all MSTF-GARCH model parameters are known, *i.e.,* the initial regimes probability $\boldsymbol{\pi}^{(0)}$, the probability tran-

sitions matrix A, and the GARCH$(1,1)$ parameters in each of the m regimes. Let $\phi \triangleq \left\{\boldsymbol{\pi}^{(0)},\ A,\ \zeta_1,...,\zeta_m,\ \alpha_1,...,\alpha_m,\ \beta_1,...,\beta_m\right\}$ be the set of parameters which specifies the model, where for a first-order process we denote $\alpha_s \triangleq \alpha_{1,s}$ and $\beta_s \triangleq \beta_{1,s}$. In practice, the model parameters ϕ are estimated from a set of clean training signals as generally done with hidden Markov models [35, 62, 64, 129] while the covariance matrix of the noise process can be estimated using the minimum statistics [72] or the minima controlled recursive averaging algorithms [38, 71]. The problem of model estimation is addressed in Section 4.5.

The spectral restoration problem is generally formulated as deriving an estimator $\hat{X}_{tk}$ for the spectral coefficients, such that the expected value of a certain distortion measure is minimized. We develop a recursive estimator for the signal's spectral coefficients and for their absolute squared values in the sense of minimum mean-square error (MMSE), and we then extend this framework to signal restoration in the sense of MMSE of the log-spectral amplitude (LSA), which is often used in speech enhancement applications, see for instance [34, 38].

Let $\mathcal{Y}^\tau = \mathcal{Y}_0^\tau \triangleq \{\mathbf{Y}_t \,|\, t = 0,...,\tau\}$ be the set of observations up to time τ. The causal MMSE estimator of the coefficients $\mathbf{X}_t$ given the noisy observations up to time t is obtained as follows:

$$E\left\{\mathbf{X}_t \,|\, \mathcal{Y}^t\right\} = \sum_{s_t} p\left(s_t \,|\, \mathcal{Y}^t\right) E\left\{\mathbf{X}_t \,|\, s_t, \mathcal{Y}^t\right\} . \tag{4.15}$$

Denote the state dependent, one-frame-ahead conditional covariance matrix of the clean signal as

$$R_{s_t}^x \triangleq E\left\{\mathbf{X}_t\mathbf{X}_t^H \,|\, s_t, \mathcal{I}^{t-1}\right\} . \tag{4.16}$$

Following the model formulation this covariance matrix is a function of $R_{s_{t-1}}^x$ and $\mathbf{X}_{t-1}$ only. However, the clean signal values are usually unavailable, nor the sequence of active states, so the evaluation of (4.16) requires the whole available observations. To overcome this problem, we assume that given current regime, past estimated conditional covariances are sufficient statistics for the conditional variance estimation [24]. Accordingly, given the set of estimated one-frame-ahead conditional variances $\hat{\Lambda}_t \triangleq \left\{\hat{\boldsymbol{\lambda}}_{t|t-1,S_t} \,|\, S_t = 1,...,m\right\}$ which manipulates the observations up to time $t-1$, we may use the following signal

estimator:

$$\hat{\mathbf{X}}_t = \sum_{s_t} p\left(s_t \,|\, \hat{\Lambda}_t, \mathbf{Y}_t\right) E\left\{\mathbf{X}_t \,|\, s_t, \hat{R}^x_{s_t}, \mathbf{Y}_t\right\}, \tag{4.17}$$

where under a Gaussian model

$$E\left\{\mathbf{X}_t \,|\, s_t, \hat{R}^x_{s_t}, \mathbf{Y}_t\right\} = \hat{R}^x_{s_t}\left(\hat{R}^x_{s_t} + R^d\right)^{-1}\mathbf{Y}_t\,. \tag{4.18}$$

Note that $\hat{R}^x_{s_t}$ is a K-by-K diagonal matrix (since $\{V_{tk}\}$ are statistically independent) with the estimated state-dependent conditional variance $\hat{\boldsymbol{\lambda}}_{t|t-1,s_t}$ on its diagonal. This state-dependent conditional variance can be recursively estimated in the MMSE sense by calculating its conditional expectation under s_t given the observation $\mathbf{Y}_{t-1}$ and the previous set of estimated conditional variances:

$$\begin{aligned}\hat{\boldsymbol{\lambda}}_{t|t-1,s_t} &\triangleq E\left\{\boldsymbol{\lambda}_{t|t-1,S_t} \,|\, s_t, \hat{\Lambda}_{t-1}, \mathbf{Y}_{t-1}; \phi\right\} \\ &= \zeta_{s_t}\mathbf{1} + \alpha_{s_t} E\left\{\mathbf{X}_{t-1}\odot\mathbf{X}^*_{t-1} \,|\, s_t, \hat{\Lambda}_{t-1}, \mathbf{Y}_{t-1}; \phi\right\} \\ &\quad + \beta_{s_t} E\left\{\boldsymbol{\lambda}_{t-1|t-2,S_{t-1}} \,|\, s_t, \hat{\Lambda}_{t-1}, \mathbf{Y}_{t-1}; \phi\right\}.\end{aligned} \tag{4.19}$$

The conditional second-order moment in (4.19), can be obtained by

$$\begin{aligned}\hat{\boldsymbol{\lambda}}_{t-1|t-1,s_t} &\triangleq E\left\{\mathbf{X}_{t-1}\odot\mathbf{X}^*_{t-1} \,|\, s_t, \hat{\Lambda}_{t-1}, \mathbf{Y}_{t-1}; \phi\right\} \\ &= \sum_{s_{t-1}} p\left(s_{t-1} \,|\, s_t, \hat{\Lambda}_{t-1}, \mathbf{Y}_{t-1}; \phi\right) E\left\{\mathbf{X}_{t-1}\odot\mathbf{X}^*_{t-1} \,|\, s_{t-1}, s_t, \hat{\Lambda}_{t-1}, \mathbf{Y}_{t-1}; \phi\right\} \\ &= \sum_{s_{t-1}} p\left(s_{t-1} \,|\, s_t, \hat{\Lambda}_{t-1}, \mathbf{Y}_{t-1}; \phi\right) E\left\{\mathbf{X}_{t-1}\odot\mathbf{X}^*_{t-1} \,|\, s_{t-1}, \hat{\boldsymbol{\lambda}}_{t-1|t-2,s_{t-1}}, \mathbf{Y}_{t-1}; \phi\right\} \\ &\triangleq \sum_{s_{t-1}} p\left(s_{t-1} \,|\, s_t, \hat{\Lambda}_{t-1}, \mathbf{Y}_{t-1}; \phi\right) \hat{\boldsymbol{\lambda}}_{t-1|t-1,s_{t-1}}\,.\end{aligned} \tag{4.20}$$

The expected one-frame-ahead conditional variance in (4.19), given the one-frame-ahead regime, can be obtained by:

$$\begin{aligned}\hat{\boldsymbol{\lambda}}_{t-1|t-2,s_t} &\triangleq E\left\{\boldsymbol{\lambda}_{t-1|t-2,S_{t-1}} \,|\, s_t, \hat{\Lambda}_{t-1}, \mathbf{Y}_{t-1}; \phi\right\} \\ &= \sum_{s_{t-1}} p\left(s_{t-1} \,|\, s_t, \hat{\Lambda}_{t-1}, \mathbf{Y}_{t-1}; \phi\right) E\left\{\boldsymbol{\lambda}_{t-1|t-2,S_{t-1}} \,|\, s_{t-1}, s_t, \hat{\Lambda}_{t-1}; \phi\right\} \\ &= \sum_{s_{t-1}} p\left(s_{t-1} \,|\, s_t, \hat{\Lambda}_{t-1}, \mathbf{Y}_{t-1}; \phi\right) \hat{\boldsymbol{\lambda}}_{t-1|t-2,s_{t-1}}\,.\end{aligned} \tag{4.21}$$

The third lines in (4.20) and in (4.21) rely on the fact that given all observations up to time $t-1$ and given the state s_{t-1}, the second-order moment of the process at that

time, and also its conditional variance, are independent of any future state. Moreover, notice that $\hat{\boldsymbol{\lambda}}_{t|t,s_t}$ and $\hat{\boldsymbol{\lambda}}_{t|t,s_{t+1}}$ in (4.20) represent the expected second-order moment of the process based on information up to time t, given the chain state at the same time, and given the next state, respectively. Similarly, $\hat{\boldsymbol{\lambda}}_{t|t-1,s_t}$ and $\hat{\boldsymbol{\lambda}}_{t|t-1,s_{t+1}}$ in (4.21) represent the expectation of the one-frame-ahead conditional variance at time t given the chain state s_t, and given the chain state at the next time step, respectively.

The MMSE estimation of the process' second-order moment $\hat{\boldsymbol{\lambda}}_{t|t,s_t}$ in (4.20) given the estimated one-frame-ahead conditional variance of the same regime $\hat{\boldsymbol{\lambda}}_{t|t-1,s_t}$ (4.21), can be obtained by

$$\begin{aligned}\hat{\boldsymbol{\lambda}}_{t|t,s_t} &= E\left\{\mathbf{X}_t \odot \mathbf{X}_t^* \,|\, s_t, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}, \mathbf{Y}_t\right\} \\ &= \hat{R}^x_{s_t}\left(\hat{R}^x_{s_t} + R^d\right)^{-1}\left[\boldsymbol{\sigma}^2 + \hat{R}^x_{s_t}\left(\hat{R}^x_{s_t} + R^d\right)^{-1}\left(\mathbf{Y}_t \odot \mathbf{Y}_t^*\right)\right], \quad s_t = 1, ..., m\,, \quad (4.22)\end{aligned}$$

similarly to the method in [24] applied to the case of a single-regime spectral GARCH. Following the notation in [24] we call (4.22) the *update* step as it updates the estimation of the signal's second-order moment at time t from its estimated one-frame-ahead conditional variance, using the new observation $\mathbf{Y}_t$. Substituting (4.20), (4.21) and (4.22) into (4.19) we obtain the *propagation* step which propagates ahead in time to obtain a conditional variance estimation at the next time, $t+1$ (assuming regime s_{t+1}), using the available information up to the current time t:

$$\hat{\boldsymbol{\lambda}}_{t+1|t,s_{t+1}} = \zeta_{s_{t+1}}\mathbf{1} + \alpha_{s_{t+1}}\hat{\boldsymbol{\lambda}}_{t|t,s_{t+1}} + \beta_{s_{t+1}}\hat{\boldsymbol{\lambda}}_{t|t-1,s_{t+1}}, \qquad s_{t+1} = 1, ..., m\,. \quad (4.23)$$

Let $\hat{\Lambda}^t \triangleq \left\{\hat{\Lambda}_0, \hat{\Lambda}_1, ..., \hat{\Lambda}_t\right\}$ be the set of the recursively estimated conditional variances up to time t, then we can manipulate all previous estimations to recursively evaluate the probability $p\left(s_{t-1} \,|\, s_t, \hat{\Lambda}^{t-1}, \mathbf{Y}_{t-1}; \phi\right)$ in (4.20) and (4.21) by

$$p\left(s_{t-1} \,|\, s_t, \hat{\Lambda}^{t-1}, \mathbf{Y}_{t-1}; \phi\right) = p\left(s_{t-1} \,|\, \hat{\Lambda}^{t-1}, \mathbf{Y}_{t-1}; \phi\right) a_{s_{t-1},s_t} / p\left(s_t \,|\, \hat{\Lambda}^{t-1}, \mathbf{Y}_{t-1}; \phi\right)\,, \quad (4.24)$$

where

$$p\left(s_t \,|\, \hat{\Lambda}^{t-1}, \mathbf{Y}_{t-1}; \phi\right) = \sum_{s_{t-1}} p\left(s_{t-1} \,|\, \hat{\Lambda}^{t-1}, \mathbf{Y}_{t-1}; \phi\right) a_{s_{t-1},s_t}\,. \quad (4.25)$$

The conditional state probability at the right of (4.25) can be obtained by

$$\begin{aligned} p\left(s_t \,|\, \hat{\Lambda}^t, \mathbf{Y}_t; \phi\right) &= \frac{b\left(\mathbf{Y}_t, s_t \,|\, \hat{\Lambda}^t; \phi\right)}{b\left(\mathbf{Y}_t \,|\, \hat{\Lambda}^t; \phi\right)} \\ &= \frac{b\left(\mathbf{Y}_t \,|\, s_t, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}\right) p\left(s_t \,|\, \hat{\Lambda}^{t-1}, \mathbf{Y}_{t-1}; \phi\right)}{\sum_{s_t} b\left(\mathbf{Y}_t | s_t, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}\right) p\left(s_t \,|\, \hat{\Lambda}^{t-1}, \mathbf{Y}_{t-1}; \phi\right)}, \end{aligned} \tag{4.26}$$

where $b\left(\cdot \,|\, \cdot\right)$ denotes a conditional density function. Specifically, $b\left(\mathbf{Y}_t \,|\, s_t, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}\right)$ is the observation conditional density which is a complex normal distribution with zero-mean and $\hat{R}^x_{s_t} + R^d$ covariance matrix,

$$b\left(\mathbf{Y}_t \,|\, s_t, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}\right) = \frac{1}{\pi^K |\hat{R}^x_{s_t} + R^d|} \exp\left\{\mathbf{Y}_t^H \left(\hat{R}^x_{s_t} + R^d\right)^{-1} \mathbf{Y}_t\right\}. \tag{4.27}$$

Computing the conditional density $b\left(\mathbf{Y}_t \,|\, s_t, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}\right)$ tends to be numerically unstable for large values of K since the diagonal values of its covariance matrix (*i.e.,* $\hat{\boldsymbol{\lambda}}_{t|t-1,s_t}$) are typically of the same order of magnitude. Therefore, $b\left(\mathbf{Y}_t \,|\, s_t, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}\right)$ tends to zero or infinity exponentially fast as K increases. It is therefore useful to recursively evaluate a normalized density $\tilde{b}\left(\mathbf{Y}_t \,|\, s_t, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}\right)$ as follows:

$$\tilde{b}\left(Y_{t,0}, ..., Y_{t\tilde{k}} \,|\, s_t, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}\right) = \frac{\tilde{b}\left(Y_{t,0}, ..., Y_{t,\tilde{k}-1} \,|\, s_t, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}\right) b\left(Y_{t,\tilde{k}} \,|\, s_t, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}\right)}{\sum_{s_t} \tilde{b}\left(Y_{t,0}, ..., Y_{t,\tilde{k}-1} \,|\, s_t, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}\right) b\left(Y_{t,\tilde{k}} \,|\, s_t, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}\right)}, \tag{4.28}$$

for $\tilde{k} = 0, ..., K-1$ and substitute it into (4.26). As can be seen from (4.26), this normalization of $b\left(\mathbf{Y}_t \,|\, s_t, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}\right)$ does not affect the value of $p\left(s_t \,|\, \hat{\Lambda}^t, \mathbf{Y}_t; \phi\right)$.

The causal one-frame-ahead conditional variance and the conditional second-order moment of the process can be obtained by

$$\begin{aligned} \hat{\boldsymbol{\lambda}}_{t|t-1} &= \sum_{s_t} p\left(s_t \,|\, \hat{\Lambda}^t, \mathbf{Y}_t\right) E\left\{\boldsymbol{\lambda}_{t|t-1,S_t} \,|\, s_t, \hat{\Lambda}_t\right\} \\ &= \sum_{s_t} p\left(s_t \,|\, \hat{\Lambda}^t, \mathbf{Y}_t\right) \hat{\boldsymbol{\lambda}}_{t|t-1,s_t} \end{aligned} \tag{4.29}$$

$$\begin{aligned} \hat{\boldsymbol{\lambda}}_{t|t} &= \sum_{s_t} p\left(s_t \,|\, \hat{\Lambda}^t, \mathbf{Y}_t\right) E\left\{\mathbf{X}_t \odot \mathbf{X}_t^* \,|\, s_t, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}, \mathbf{Y}_t\right\} \\ &= \sum_{s_t} p\left(s_t \,|\, \hat{\Lambda}^t, \mathbf{Y}_t\right) \hat{\boldsymbol{\lambda}}_{t|t,s_t}, \end{aligned} \tag{4.30}$$

while a state smoothing (*i.e.*, noncausal state probability estimation) for the path-dependent MSTF-GARCH model has been derived in [130] and may be employed for noncausal estimation.

The causal recursive MMSE signal restoration algorithm, presented in (4.17) to (4.26), has a compact vector form with respect to the regimes vector. Let $\mathbf{s}_t \triangleq [S_t = 1, ..., S_t = m]'$ be the regimes vector at time t, let

$$\boldsymbol{\rho}_t(\mathbf{s}_\tau) \triangleq \left[p\left(S_\tau = 1 \,|\, \hat{\Lambda}^t, \mathbf{Y}_t\right), ..., p\left(S_\tau = m \,|\, \hat{\Lambda}^t, \mathbf{Y}_t\right)\right]' \tag{4.31}$$

be the probabilities of the regimes vector $\mathbf{s}_\tau$, conditioned on all observations up to frame t. Let C_t be a regimes probability matrix at time t conditioned on the next regime and all available observations up to time t, *i.e.*, $c_{t,ij} = p\left(S_t = i \,|\, S_{t+1} = j,\, \hat{\Lambda}^t, \mathbf{Y}_t\right), \quad i, j = 1, ..., m$. Let $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ represent the vectors of the m regimes' GARCH parameters, *i.e.*, $\boldsymbol{\alpha} \triangleq [\alpha_1, ..., \alpha_m]'$ and $\boldsymbol{\beta} \triangleq [\beta_1, ..., \beta_m]'$. Let $\hat{\boldsymbol{\lambda}}_{tk|\tau_1,\mathbf{s}_{\tau_2}} \triangleq \left[\hat{\lambda}_{tk|\tau_1,S_{\tau_2}=1}, ..., \hat{\lambda}_{tk|\tau_1,S_{\tau_2}=m}\right]'$ be an $m \times 1$ vector of the kth index estimated conditional variances based on observations up to time τ_1, and the corresponding m regimes vector $\mathbf{s}_{\tau_2}$. Denote by $\mathbf{a}^{(i)}$ and $\mathbf{c}_t^{(i)}$ the ith column of matrices A and C_t, respectively, and let $(\div)$ denote a term-by-term division of two vectors. A step-by-step vector form of the causal signal estimation procedure is described in Table 4.1.

The algorithm, summarized in Table 4.1, estimates both the spectral coefficients and their conditional variance in the MMSE sense. A more general signal enhancement problem is formulated as minimization of the following distortion measure:

$$E\left\{|f\left(X_{tk}\right) - f(\hat{X}_{tk})|^2 \,|\, \mathcal{Y}^t\right\}, \tag{4.32}$$

where $f(X)$ is a Borel integrable function. The estimator can be found from

$$f\left(\hat{X}_{tk}\right) = E\left\{f\left(X_{tk}\right) \,|\, \mathcal{Y}^t\right\}, \tag{4.33}$$

where

$$E\left\{f\left(X_{tk}\right) \,|\, \mathcal{Y}^t\right\} = \sum_{s_t} p\left(S_t = s_t \,|\, \mathcal{Y}^t\right) E\left\{f\left(X_{tk}\right) \,|\, s_t, \mathcal{Y}^t\right\}. \tag{4.34}$$

The log-spectral amplitude MMSE estimator, obtained by substituting $f\left(X\right) = \log|X|$ into (4.33), is of particular importance in speech enhancement applications, see for instance [34, 38]. The LSA estimator [34] is given by

Table 4.1: Vector form of the recursive MSTF-GARCH signal estimation

Initialization:

$$\boldsymbol{\rho}_{-1}(\mathbf{s}_0) = \boldsymbol{\pi}$$

$$\hat{\boldsymbol{\lambda}}_{-1,k|-2,\mathbf{s}_0} = \hat{\boldsymbol{\lambda}}_{-1,k|-1,\mathbf{s}_0} = \mathbf{0}_{m\times 1}, \qquad k = 0, ..., K-1$$

for $t = 0, ..., T-1$

$$\hat{\boldsymbol{\lambda}}_{tk|t-1,\mathbf{s}_t} = \boldsymbol{\zeta} + \boldsymbol{\alpha} \odot \hat{\boldsymbol{\lambda}}_{t-1,k|t-1,\mathbf{s}_t} + \boldsymbol{\beta} \odot \hat{\boldsymbol{\lambda}}_{t-1,k|t-2,\mathbf{s}_t}, \qquad k = 0, ..., K-1$$

$$\hat{\boldsymbol{\lambda}}_{tk|t,\mathbf{s}_t} = \hat{\boldsymbol{\lambda}}_{tk|t-1,\mathbf{s}_t} \odot \left[\sigma_k^2\mathbf{1} + \left(\hat{\boldsymbol{\lambda}}_{tk|t-1,\mathbf{s}_t} \cdot |Y_{tk}|^2\right)(\div)\left(\hat{\boldsymbol{\lambda}}_{tk|t-1,\mathbf{s}_t} + \sigma_k^2\mathbf{1}\right)\right] (\div)\left(\hat{\boldsymbol{\lambda}}_{tk|t-1,\mathbf{s}_t} + \sigma_k^2\mathbf{1}\right), \qquad k = 0, ..., K-1$$

$$b\left(\mathbf{Y}_t|s_t, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}\right) = \pi^{-K}|\hat{R}^x_{s_t} + R_d|^{-1}\exp\left\{-\mathbf{Y}_t^H\left(\hat{R}^x_{s_t} + R_d\right)^{-1}\mathbf{Y}_t\right\}, \quad s_t = 1, ..., m$$

$$B_t \triangleq \mathrm{diag}\left\{\mathbf{b}\left(\mathbf{Y}_t|\mathbf{s}_t, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}\right)\right\}$$

$$\boldsymbol{\rho}_t(\mathbf{s}_t) = B_t\ \boldsymbol{\rho}_{t-1}(\mathbf{s}_t)\left[\mathbf{1}'B_t\ \boldsymbol{\rho}_{t-1}(\mathbf{s}_t)\right]^{-1}$$

$$\boldsymbol{\rho}_t(\mathbf{s}_{t+1}) = A'\boldsymbol{\rho}_t(\mathbf{s}_t)$$

$$for\ i = 1, ..., m:\ \mathbf{c}_t^{(i)} = \mathbf{a}^{(i)} \odot \boldsymbol{\rho}_t(\mathbf{s}_t)/\rho_t(s_{t+1} = i)$$

$$\hat{\boldsymbol{\lambda}}_{tk|t,\mathbf{s}_{t+1}} = C_t'\ \hat{\boldsymbol{\lambda}}_{tk|t,\mathbf{s}_t}, \qquad k = 0, ..., K-1$$

$$\hat{\boldsymbol{\lambda}}_{tk|t-1,\mathbf{s}_{t+1}} = C_t'\ \hat{\boldsymbol{\lambda}}_{tk|t-1,\mathbf{s}_t}, \qquad k = 0, ..., K-1$$

$$\hat{\mathbf{X}}_{t|t,s_t} = \hat{R}^x_{s_t}\left(\hat{R}^x_{s_t} + R^d\right)^{-1}\mathbf{Y}_t$$

$$\hat{X}_{tk} = \boldsymbol{\rho}_t'(\mathbf{s}_t)\hat{X}_{tk|t,s_t}, \qquad k = 0, ..., K-1$$

$$\begin{aligned}\widehat{|X_{tk}|} &= \exp\left(E\left\{\log|X_{tk}|\,|\,\lambda_{tk}, Y_{tk}\right\}\right)\\ &= G\left(\xi_{tk}, \vartheta_{tk}\right)|Y_{tk}|,\end{aligned} \tag{4.35}$$

where

$$\xi_{tk} \triangleq \frac{\lambda_{tk}}{\sigma^2_{tk}}, \qquad \gamma_{tk} \triangleq \frac{|Y_{tk}|^2}{\sigma^2_{tk}}, \qquad \vartheta_{tk} \triangleq \frac{\gamma_{tk}\xi_{tk}}{1+\xi_{tk}} \tag{4.36}$$

and

$$G(\xi, \vartheta) = \frac{\xi}{1+\xi}\exp\left(\frac{1}{2}\int_{\vartheta}^{\infty}\frac{e^{-t}}{t}dt\right). \tag{4.37}$$

ξ_{tk} and γ_{tk} represent the *a priori* and *a posteriori* SNRs respectively [33].

By substituting (4.35) into (4.34), and combining the result with the phase of the noisy signal [34], we obtain the spectral coefficient estimator in the MMSE-LSA sense

$$\hat{X}_{tk} = Y_{tk} \prod_{s_t} G\left(\hat{\xi}_{tk,s_t}, \hat{\vartheta}_{tk,s_t}\right)^{p\left(s_t|\mathcal{Y}^t\right)}, \tag{4.38}$$

where

$$\hat{\xi}_{tk,s_t} \triangleq \frac{\hat{\lambda}_{tk|t,s_t}}{\sigma^2_{tk}}, \qquad \hat{\vartheta}_{tk,s_t} \triangleq \frac{\gamma_{tk}\hat{\xi}_{tk,s_t}}{1+\hat{\xi}_{tk,s_t}}, \tag{4.39}$$

and $p\left(s_t|\mathcal{Y}^t\right)$ is recursively estimated using (4.26).

4.4 Estimation efficiency

In this section we analyze the mean-square error of a one step ahead MMSE estimation using the proposed recursive algorithm. The recursive formulation of the MSTF-GARCH yields an accumulated error in the estimation of the variance and the signal. However, for each regime and in each frame the algorithm evaluates the conditional variance as a weighted sum of previous estimated conditional variances and squared absolute values (4.20), (4.21). These weights are proportional to the conditional densities $b\left(\mathbf{Y}_t \,|\, s_t, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}\right)$ in (4.26). Consequently, an over estimation of the conditional variance on a specific frame can be followed in the algorithm by giving a high probability (*i.e.,* higher weight) to a regime with small parameters which compensates the previous over estimation. Similarly, an under estimation of the conditional variance can be compensated by giving a high probability to a regime with large parameters.

Assume that the process is observed perfectly (without noise) up to time $t-1$ and that the regime path is known up to that time. Then, $\boldsymbol{\lambda}_{t-1|t-2,s_{t-1}}$ can be calculated by (4.5). Following Ephraim and Merhav [117] which derive bounds for the MSE of a composite source signal estimation, we assume that (i) the Markov chain is stationary and the necessary and sufficient condition for a bounded variance is satisfied; (ii) $\hat{\boldsymbol{\lambda}}_{t|t,s_t}$ is square integrable with respect to $b\left(\mathbf{Y}_t \,|\, \boldsymbol{\lambda}_{t|t-1,s_t}\right)$ and $b\left(\mathbf{Y}_t \,|\, \boldsymbol{\lambda}_{t|t-1,\tilde{s}_t}\right)$; and (iii) the regime transition probabilities are positive, *i.e.,* $a_{ij} \geq a_{\min} > 0 \;\; \forall i,j = 1,...,m$.

The one-step-ahead MMSE estimator (4.17) is unbiased in the sense that $E\left\{\hat{\mathbf{X}}_t\right\} = E\left\{\mathbf{X}_t\right\}$, and following [117] we obtain an upper bound for the variance of the one-step-ahead estimation error, assuming that the process is observed with an additive, independent stationary noise. The one-step-ahead MSE is given by

$$\overline{e_t^2} \triangleq \frac{1}{K}\mathrm{tr}E\left\{\left(\mathbf{X}_t-\hat{\mathbf{X}}_t\right)\left(\mathbf{X}_t-\hat{\mathbf{X}}_t\right)^H\right\}, \tag{4.40}$$

where the signal estimator $\hat{\mathbf{X}}_t$ follows

$$\begin{aligned}\hat{\mathbf{X}}_t &= E\left\{\mathbf{X}_t \,|\, \mathcal{I}^{t-1}, \mathbf{Y}_t\right\} \\ &= E\left\{\mathbf{X}_t \,|\, s_{t-1}, \boldsymbol{\lambda}_{t-1|t-2,s_{t-1}}, \mathbf{X}_{t-1}, \mathbf{Y}_t\right\} \\ &= E\left\{\mathbf{X}_t \,|\, \Lambda_t, \mathbf{Y}_t\right\}. \end{aligned} \tag{4.41}$$

Under the above assumptions the MSE can be written as [117, eq. (13) – (17)]

$$\overline{e_t^2} = \overline{\mu_t^2} + \overline{\eta_t^2}, \tag{4.42}$$

where

$$\begin{aligned}\overline{\mu_t^2} &\triangleq \frac{1}{K}\mathrm{tr}E\left\{\mathrm{cov}\left(\mathbf{X}_t \,|\, \Lambda_t, s_t, \mathbf{Y}_t\right)\right\} \\ &= \frac{1}{K}\mathrm{tr}E\left\{\mathrm{cov}\left(\mathbf{X}_t \,|\, \boldsymbol{\lambda}_{t|t-1,s_t}, s_t, \mathbf{Y}_t\right)\right\} \end{aligned} \tag{4.43}$$

$$\overline{\eta_t^2} \triangleq \frac{1}{2}\sum_{s_t\neq\tilde{s}_t} E\left\{p\left(s_t \,|\, s_{t-1}, \Lambda_t, \mathbf{Y}_t\right) p\left(\tilde{s}_t \,|\, s_{t-1}, \Lambda_t, \mathbf{Y}_t\right) g\left(s_t, \tilde{s}_t, \Lambda_t, \mathbf{Y}_t\right)\right\}, \tag{4.44}$$

and

$$\begin{aligned} g\left(s_t, \tilde{s}_t, \Lambda_t, \mathbf{Y}_t\right) &\triangleq \frac{1}{K}\mathrm{tr}\left\{\left(\hat{\mathbf{X}}_{t|t,s_t}-\hat{\mathbf{X}}_{t|t,\tilde{s}_t}\right)\left(\hat{\mathbf{X}}_{t|t,s_t}-\hat{\mathbf{X}}_{t|t,\tilde{s}_t}\right)^H\right\} \\ &= g\left(s_t, \tilde{s}_t, \boldsymbol{\lambda}_{t|t-1,s_t}, \boldsymbol{\lambda}_{t|t-1,\tilde{s}_t}, \mathbf{Y}_t\right). \end{aligned} \tag{4.45}$$

The state probabilities in (4.44) can be evaluated using (4.26):

$$p\left(s_t \,|\, s_{t-1}, \Lambda_t, \mathbf{Y}_t\right) = \frac{b\left(\mathbf{Y}_t \,|\, s_t, \boldsymbol{\lambda}_{t|t-1,s_t}\right) a_{s_{t-1},s_t}}{\sum_{s_t} b\left(\mathbf{Y}_t \,|\, s_t, \boldsymbol{\lambda}_{t|t-1,s_t}\right) a_{s_{t-1},s_t}}, \tag{4.46}$$

and the signal estimate given the state s_t is given by

$$\hat{\mathbf{X}}_{t|t,s_t} = W_{s_t}\mathbf{Y}_t, \tag{4.47}$$

where W_{s_t} is the conditional Wiener filter: $W_{s_t} \triangleq R_{s_t}^x\left(R_{s_t}^x + R^d\right)^{-1}$. Substituting (4.47) into (4.45), we have

$$\begin{aligned} g\left(s_t, \tilde{s}_t, \Lambda_t, \mathbf{Y}_t\right) &= \frac{1}{K}\mathbf{Y}_t^H \left(W_{s_t} - W_{\tilde{s}_t}\right)^H \left(W_{s_t} - W_{\tilde{s}_t}\right)\mathbf{Y}_t \\ &\triangleq \frac{1}{K}\mathbf{Y}_t^H W_{\tilde{s}_t s_t}^2 \mathbf{Y}_t \,. \end{aligned} \tag{4.48}$$

The one-step-ahead MSE, $\overline{e_t^2}$, is decomposed into two positive terms, $\overline{\mu_t^2}$ and $\overline{\eta_t^2}$. The first is the MSE of the estimator $\hat{\mathbf{X}}_{t|t,s_t}$ which relies on knowing the true regime at time t, and therefore it is the optimal estimator in the MMSE sense. This term is evaluated by substituting (4.43) into (4.47):

$$\overline{\mu_t^2} = \frac{1}{K}\mathrm{tr}\sum_{s_t} a_{s_{t-1}s_t} W_{s_t} R^d \,. \tag{4.49}$$

The second term, $\overline{\eta_t^2}$, is a weighted sum of cross error terms which depend on pairs of the process regimes. This term is difficult to evaluate, but it is upper bounded by [117, eq. (18) and (23)]

$$\overline{\eta_t^2} \leq \frac{1}{2}\sum_{s_t \neq \tilde{s}_t} a_{\min}^{-2}\left(I_{s_t}\left(\tilde{s}_t\right) + I_{\tilde{s}_t}\left(s_t\right)\right) \,, \tag{4.50}$$

where

$$I_{s_t}\left(\tilde{s}_t\right) \leq \frac{1}{K}\sum_{s_t \neq \tilde{s}_t} \mathrm{tr}\left\{W_{s_t\tilde{s}_t}^2 Q_{\tilde{s}_t}\right\}\left(\left|R_\lambda\left(s_t, \tilde{s}_t\right)\right| \cdot \left|Q_{s_t}\right|^{-\lambda} \cdot \left|Q_{\tilde{s}_t}\right|^{\lambda-1} + \frac{\mathrm{tr}\left\{W_{s_t\tilde{s}_t}^2 R_\lambda\left(s_t, \tilde{s}_t\right)\right\}}{\mathrm{tr}\left\{W_{s_t\tilde{s}_t}^2 Q_{\tilde{s}_t}\right\}}\right) \tag{4.51}$$

with $\lambda > 0$ [117, eq. (31) to (39) and (54) to (60)], Q_{s_t} denotes the covariance matrix of the noisy signal given the regime s_t, and $R_\lambda\left(s_t, \tilde{s}_t\right)$ is defined by

$$R_\lambda\left(s_t, \tilde{s}_t\right) \triangleq \left[\lambda Q_{s_t}^{-1} + \left(1-\lambda\right)Q_{\tilde{s}_t}^{-1}\right]^{-1} \,. \tag{4.52}$$

In the derivation of (4.51) it is assumed that $R_\lambda\left(s_t, \tilde{s}_t\right)$ is positive definite [117]. Since Q_{s_t} is a diagonal matrix with positive eigenvalues, $R_\lambda\left(s_t, \tilde{s}_t\right)$ is positive definite for any $0 < \lambda < 1$. Substituting (4.51) into (4.50) and using the diagonality of the covariance matrices, we obtain an upper bound for the cross error term

$$\overline{\eta_t^2} \leq \frac{1}{a_{\min}^2 K}\sum_{s_t \neq \tilde{s}_t}\left(\mathrm{tr}\left\{W_{s_t\tilde{s}_t}^2 Q_{\tilde{s}_t}\right\} \cdot \left|R_\lambda\left(s_t, \tilde{s}_t\right)\right| \cdot \left|Q_{s_t}\right|^{-\lambda} \cdot \left|Q_{\tilde{s}_t}\right|^{\lambda-1} + \mathrm{tr}\left\{W_{s_t\tilde{s}_t}^2 R_\lambda\left(s_t, \tilde{s}_t\right)\right\}\right) \tag{4.53}$$

for $0 < \lambda < 1$. It is worthwhile noting that our MSE analysis follows the analysis in [117] but, the latter deals with a memoryless regime-switching process and a Toeplitz covariance matrix, whereas in our case both assumptions do not hold.

4.5 Model estimation

In this section we address the problem of estimating the model parameters $\phi \triangleq \{\boldsymbol{\pi}^{(0)}, A, \zeta_1, ..., \zeta_m, \alpha_1, ..., \alpha_m, \beta_1, ..., \beta_m\}$. The ML estimation approach is commonly used for estimating the parameters of GARCH models (*e.g.*, [1, 5, 7]) and also for estimating the transition probability matrices (*e.g.,* [129]). The model parameters are estimated from a training data set of N clean signals of lengths T_n ,$n = 1, ..., N$. Let $\left\{\mathbf{X}_t^{(n)}\right\}$ denote the spectral coefficients of the nth clean training signal and let $\mathcal{X}^{\tau,(n)} \triangleq \{\mathbf{X}_t^{(n)} \,|\, t = 0, ..., \tau\}$. The conditional distribution of the vector $\mathbf{X}_t^{(n)}$ given its past observations is a mixing of zero mean Gaussian vectors with diagonal covariance matrices $\hat{R}_{s_t}^{x,(n)}$:

$$b\left(\mathbf{X}_t^{(n)} \,|\, \mathcal{X}^{t-1,(n)}\right) = \sum_{s_t} p\left(s_t \,|\, \mathcal{X}^{t-1,(n)}\right) b\left(\mathbf{X}_t^{(n)} \,|\, s_t, \hat{R}_{s_t}^{x,(n)}\right) . \tag{4.54}$$

Given a set of model parameters ϕ, the diagonal covariance matrix of the density $b\left(\mathbf{X}_t^{(n)} \,|\, s_t, R_{s_t}^{x,(n)}\right)$ can be recursively estimated by using the estimation algorithm introduced in Section 4.3, where the signal observations are known in this case. Assuming that the process is asymptotically wide-sense stationary, and that the training sequences are sufficiently large, the initial state probabilities, $\boldsymbol{\pi}^{(0)}$, and the initial conditional variance, $\boldsymbol{\lambda}_{0|-1,s_0}$, have negligible contribution to the total likelihood. Therefore it is convenient to choose in the following optimization problem the stationary values as the initial values, *i.e.*, $\hat{\boldsymbol{\lambda}}_{0k|-1,\mathbf{s}_0} = \hat{\Phi}\hat{\boldsymbol{\zeta}}$ as the initial conditional variances, and $\boldsymbol{\pi}^{(0)} = \hat{\boldsymbol{\pi}}$ as the initial state probabilities.

The conditional log-likelihood of the training set is given by

$$\mathcal{L}(\phi) = \sum_n \sum_{t=0}^{T_n - 1} \log b\left(\mathbf{X}_t^{(n)} \,|\, \mathcal{X}^{t-1,(n)}; \phi\right) . \tag{4.55}$$

Using the constraints in (4.6) and imposing $\hat{A}$ to be a transition probability matrix, the ML estimates of the model parameters ϕ can be obtained by solving the following nonlinear constrained optimization problem:

$$\max_{\hat{\phi}} \mathcal{L}(\phi)$$

$$s.t. \;\; \hat{\zeta}_i{>}0, \quad \hat{\alpha}_i \geq 0, \quad \hat{\beta}_i \geq 0, \quad \sum_{j=1}^{M} \hat{a}_{ij} = 1 \quad \forall i \in \{1, ..., m\} \; . \tag{4.56}$$

For a given parameters set $\hat{\phi}$, the sequence of state dependent conditional variances $\left\{\hat{\Lambda}_t\right\}$ can be evaluated recursively according to the method described in Section 4.3 and so is the set of conditional state probabilities $p\left(s_t \mid \mathcal{X}^{t-1}\right)$. The conditional log-likelihood (4.55) can then numerically maximized under the linear constrains of (4.56) as specified in [6, 7] or by using sequential quadratic programming [131, 132]. The computational complexity required for the model estimation is much higher than that required for a single-regime GARCH model since m^2 parameters are to be estimated for the transition probabilities matrix and in addition $3m$ GARCH parameters are to be evaluated. However, using the Markov-switching model, the optimization problem needs to be solved only once, prior to the restoration procedure. It is well known that the optimal set of parameters, ϕ, is not necessarily unique in a Markovian model [129] and in addition, the numerical optimization solution may only guarantee a local maxima of the likelihood function. However, the flexibility of the model enables better results than that achievable with a single-regime GARCH model [7,8]. This is also shown in our simulation results, both for MSTF-GARCH processes and for speech signals.

4.6 Experimental results

In this section we demonstrate the performance of the proposed algorithm when applied to restoration of noisy MSTF-GARCH signals, and to estimation of conditional variances and squared absolute values of speech signals in the STFT domain.

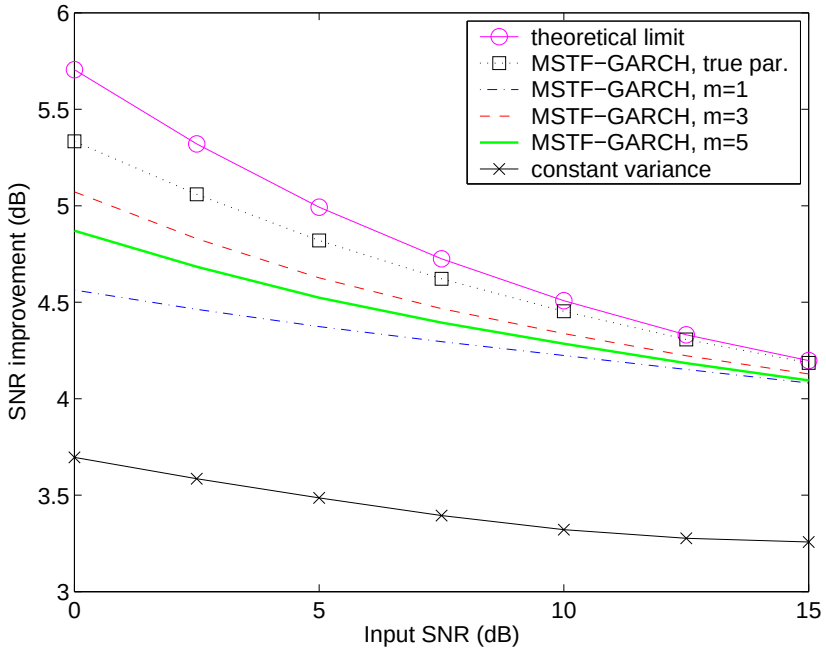
4.6.1 MSTF-GARCH signals

The proposed model estimation and signal restoration algorithm has been applied to MSTF-GARCH models of 3 and 5 regimes, degraded by additive independent white noise with 0 to 15 dB input signal-to-noise ratio (SNR). For each state space ($m = 3, 5$), a set of 20 stationary models have been simulated with uniformly distributed parameters on the

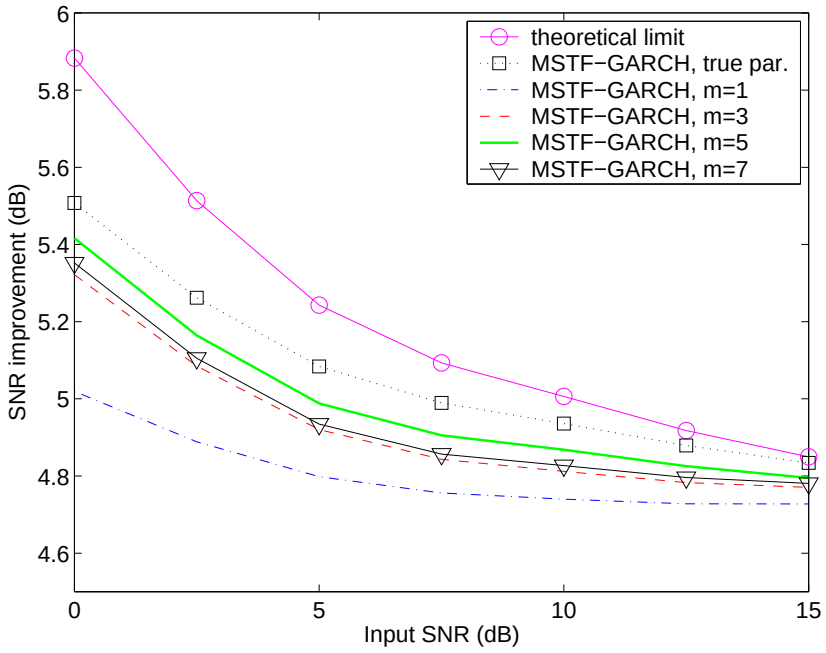
interval $(0, 1]$. For each model, the parameters, ϕ, are estimated from a set of 10 training signals, each of time length $T = 100$ and dimension $K = 100$. The estimated parameters are employed for restoration of a set of test signals containing 20 noisy signals of the same size, and basically 4 types of estimated variances are compared by incorporating them into the signal's recursive MMSE estimator of (4.17) and (4.18). The "*theoretical limit*" is referred to as the estimator which exploits the true conditional variances, $\boldsymbol{\lambda}_{t|t-1,s_t}$, of the simulated process. This estimator is the optimal estimator in the MMSE sense and its performance is compared with those of the recursive estimators. The "*MSTF-GARCH, true model*" is referred to as the recursive signal estimator, described in Section 4.3, which manipulates the true parameters set, ϕ, and the "*MSTF-GARCH, $m = i$*" estimator employs a set of estimated parameters, $\hat{\phi}$, assuming that the model has i regimes. For the "*MSTF-GARCH, $m = i$*" estimator, the set of parameters, $\hat{\phi}$, is estimated using the ML approach as described in Section 4.5. The performance of our algorithm is also compared with that of an estimator that assumes a "*constant variance*" process. For that estimator (only), the vector of "stationary" variances, are evaluated for each noisy signal from the corresponding clean signal.

Figure 4.1 (a) shows the SNR improvement obtained by using the different estimators, when applied to 3-state MSTF-GARCH signals. It can be seen that even when assuming a small number of regimes, still the MSTF-GARCH estimator outperforms the "*constant variance*" estimator, and the results achieved by assuming 3 or 5 regimes are comparable to those obtained by using the true model parameters. Figure 4.1 (b) shows estimation results for 5-state MSTF-GARCH processes, under the assumption of 1, 3, 5 or 7 regimes. The estimation performances improve with the increase of the number of assumed regimes, but using a larger number of regimes than the true number (*e.g.*, 7 instead of 5 or 5 instead of 3) yields less accurate results.

The time-varying behavior of the recursive estimator is demonstrated for a 5-state MSTF-GARCH signal degraded by additive white noise with 5 dB SNR. Figure 4.2 shows trace of the instantaneous output SNR for each time frame, obtained by the optimal estimator, the recursive estimators with presumable 1 or 5 regimes (*i.e.,* "*MSTF-GARCH, $m = 1, 5$*") and a "*constant variance*" estimator. The varying volatility of the process implies time-varying performances for all those estimators. Nevertheless, under the as-



(a)



(b)

Figure 4.1: SNR improvements obtained by using different MSTF-GARCH based estimators when applied to (a) 3-states MSTF-GARCH signals and (b) 5-state MSTF-GARCH signals. MSTF-GARCH models with various number of regimes are considered and compared with the true MSTF-GARCH parameters, the theoretical limit and a constant variance estimator.

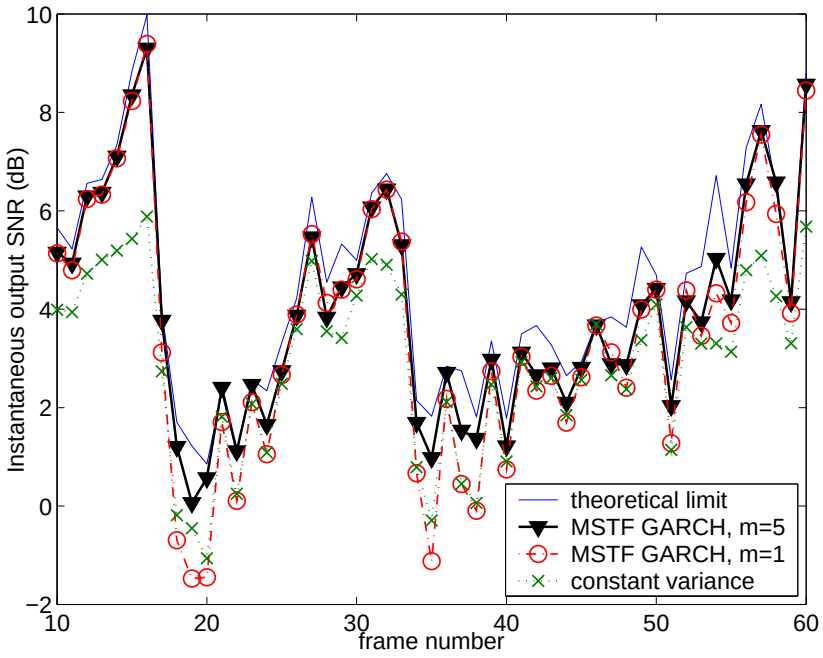


Figure 4.2: Trace of instantaneous output SNR achieved by the proposed algorithm when applied to a realization of a 5-state MSTF-GARCH process degraded by additive white noise with 5 dB SNR, and restored by an MSTF-GARCH estimator, assuming 1 and 5 states.

sumption of 5 regimes our recursive estimator follows the optimal estimator with a relatively small degradation in performance. The single-regime estimator yields comparable results as the 5-regimes estimator for frames with large input SNR. However, for frames with low input SNR the results obtained by the single-regime estimator are comparable to those obtained by the "*constant variance*" estimator.

4.6.2 Speech signals

The idea of using different states for the enhancement of speech signals was first introduced by Drucker [28]. He assumed five categories of speech signals, comprising fricatives, stops, vowels, glides, and nasals. The application of HMMs for speech enhancement requires a higher number of states [35, 62] since these models allow only a single density, or a finite set of mixture-densities, for the spectral coefficients in each state. The GARCH-based models allow continuous values of conditional variances with possible transients resulting from switching states. Hence, a small number of states may be sufficient for the representation of the coefficients' second-order moments. Furthermore, the dynamic of the

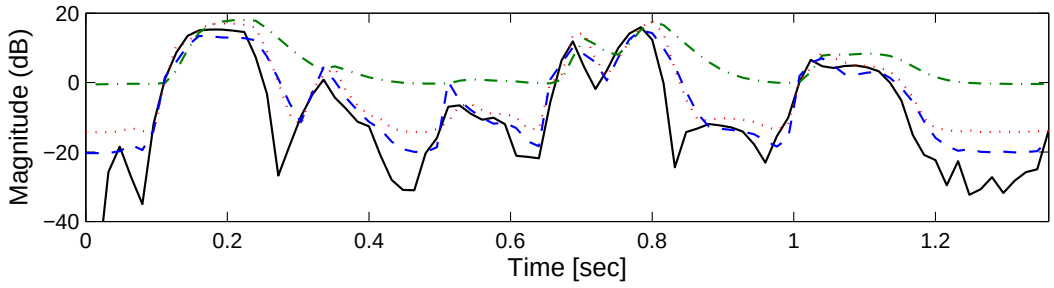
spectral coefficients is frequency dependent. Therefore, we assume different parameters in different sub-bands.

The speech signals used in our evaluation are taken from the TIMIT database. The training set includes 10 different utterances from 10 different speakers, half male and half female. The speech signals are sampled at 8 kHz and normalized to the same energy. Transformation into the STFT domain is obtained by using half overlapping Hamming analysis window of 32 millisecond length. We consider 1,3 and 5-state MSTF-GARCH models for the speech signals and estimate the one-frame-ahead conditional variance for test speech signals, not on the training set. Figure 4.3 shows typical estimates of the one-frame-ahead conditional variance, $\hat{\lambda}_{tk|t-1}$, at frequencies of 1, 2 and 3 kHz, using the different MSTF-GARCH models and assuming independent model parameters in each frequency sub-band. The estimated conditional variances are compared with the clean signal's squared absolute value $|X_{tk}|^2$. It can be seen that by increasing the number of regimes, the conditional variance yields a better prediction of the squared absolute value of the signal. Moreover, it can be seen that the conditional variance estimated by a single-regime model is smoother than that estimated based on a multi-regime model, and the latter better tracks rapid changes in the signal's energy with possible switching of regimes. During the first few frames, the speech signal is absent and thus, as long as the squared absolute value is below the minimum variance allowed by the model, the predicted variances are determined by the model threshold. However, the predicted variances converge to the absolute squared value as soon as the latter exceeds this threshold. Larger number of states may allow better representation of the conditional variance in different magnitude ranges and different volatilities, at the expense of greater computational complexity.

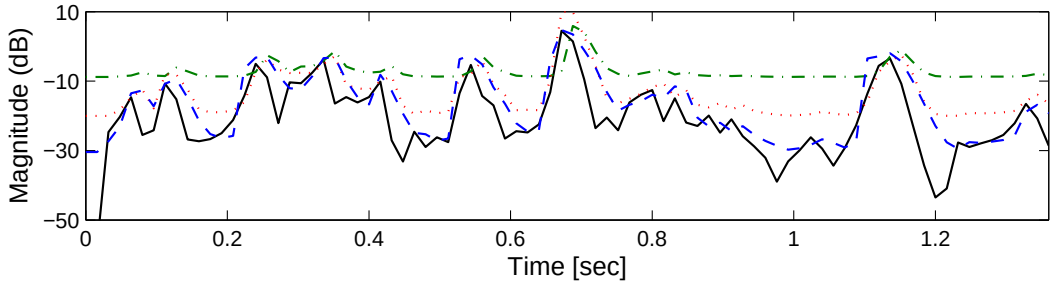
Many speech enhancement algorithms employ the decision-directed approach for the speech spectral variance estimation [33, 70]. Accordingly,

$$\hat{\lambda}_{tk}^{DD} = \max\left\{\bar{\alpha}\left|\hat{X}_{t-1,k}\right|^2 + (1-\bar{\alpha})\left(|Y_{tk}|^2 - \sigma_k^2\right)\ ,\ \xi_{\min}\sigma_k^2\right\}\ , \tag{4.57}$$

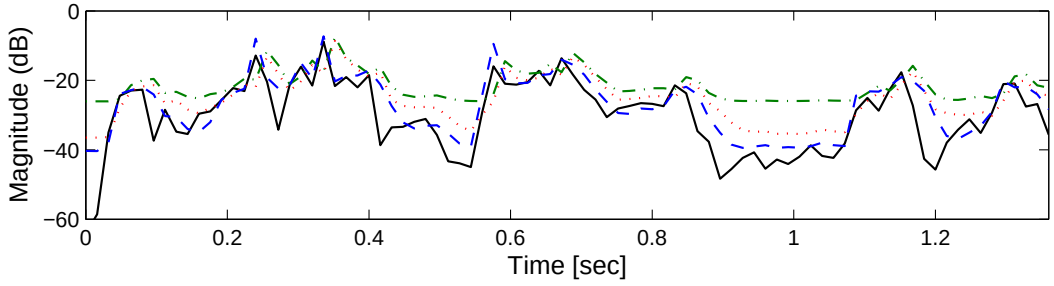
where $\bar{\alpha}$ $(0 \leq \bar{\alpha} \leq 1)$ is a weighting factor that controls the trade-off between noise reduction and transient distortion introduced into the signal. A larger value of $\bar{\alpha}$ results in a greater reduction of the musical noise phenomena, but at the expense of attenuated speech onsets and audible modifications of transient components. The parameter $\xi_{\min}$ is



(a)

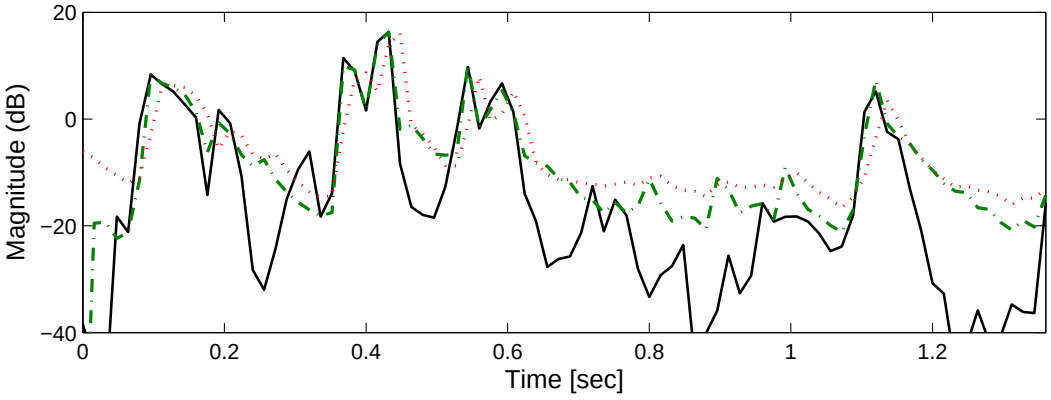


(b)

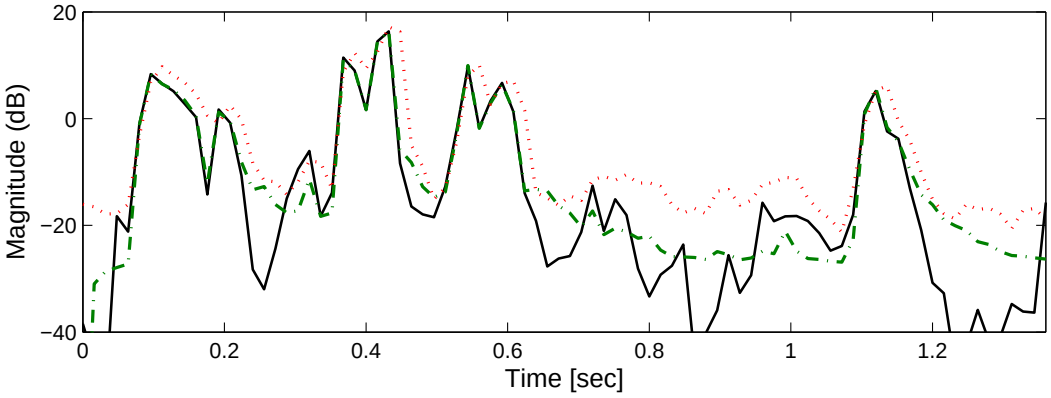


(c)

Figure 4.3: Typical traces of one-frame-ahead conditional variance estimates for speech signals at frequencies (a) 1 kHz, (b) 2 kHz and (c) 3 kHz. The conditional variances are estimated by MSTF-GARCH models of single-state (dashed-dotted line), 3 states (dotted line) and 5 states (dashed line), and compared with the clean signal's squared absolute value (solid line).



(a)



(b)

Figure 4.4: Typical traces of estimated squared absolute values for speech signal at frequency of 2 kHz. The variances are estimated by a 5-state MSTF-GARCH model (dashed-dotted line), decision-directed approach (dotted line) and compared with the clean signal's squared absolute value (solid line). The SNRs are (a) 0 dB and (b) 10 dB.

a lower bound on the a priori SNR.

The GARCH modeling enables an analytical derivation of the decision-directed estimator [69]. Considering the degenerated case of a single-state and a single-frequency ARCH(1) model (*i.e.,* $\beta = 0$), the update step (4.22) can be written as

$$\hat{\lambda}_{tk|t} = \bar{\alpha}_{tk}\hat{\lambda}_{tk|t-1} + (1 - \bar{\alpha}_{tk})\left(|Y_{tk}|^2 - \sigma_k^2\right) \tag{4.58}$$

with

$$\bar{\alpha}_{tk} \triangleq 1 - \frac{\hat{\lambda}_{tk|t-1}^2}{\left(\hat{\lambda}_{tk|t-1} + \sigma_k^2\right)^2}, \tag{4.59}$$

and $0 < \bar{\alpha}_{tk} < 1$. Substituting the propagation step for $\hat{\lambda}_{tk|t-1}$ (4.23) into (4.58) with $\alpha = 1$, we obtain

$$\hat{\lambda}_{tk|t} = \bar{\alpha}_{tk}E\left\{|X_{t-1,k}|^2 \mid \mathcal{Y}^{t-1}\right\} + (1 - \bar{\alpha}_{tk})\left(|Y_{tk}|^2 - \sigma_k^2\right) + \bar{\alpha}_{tk}\zeta\,. \tag{4.60}$$

For $\zeta << E\left\{|X_{t-1,k}|^2 \mid \mathcal{Y}^{t-1}\right\}$, (4.60) is similar to the decision-directed variance estimation (4.57) with $\bar{\alpha}_{tk} \equiv \bar{\alpha}$ and where $E\left\{|X_{t-1,k}|^2 \mid \mathcal{Y}^{t-1}\right\}$ holds for $\left|\hat{X}_{t-1,k}\right|^2$ which is the squared absolute value of the spectral coefficient estimate based on the observations $\mathcal{Y}^{t-1}$. Accordingly, the degenerated ARCH-based variance estimation with $\alpha = 1$ and low valued ζ is closely related to the decision-directed estimator with a time-varying frequency-dependent weighting factor $\bar{\alpha}_{tk}$. However, the GARCH (and ARCH) modeling approach manipulates the spectral variance as a random process, whereas the decision-directed approach assumes the spectral variance is a parameter which is heuristically evaluated. In addition, the decision-directed approach thresholds the estimated variance to be larger than $\xi_{\min}\sigma_k^2$ while in the GARCH modeling, the lower bound is inherently incorporated into the variance estimation. Since $\hat{\lambda}_{tk|t-1} > \zeta$, from (4.22) we obtain the following lower bound

$$\hat{\lambda}_{tk|t} > \frac{\zeta}{\zeta + \sigma_k^2}\left(\sigma_k^2 + \frac{\zeta}{\zeta + \sigma_k^2}|Y_{tk}|^2\right) > 0\,. \tag{4.61}$$

Modeling the spectral coefficients as an MSTF-GARCH allows further flexibility for the variance estimation. Figure 4.4 demonstrates the estimated squared absolute values of a speech signal corrupted by a white Gaussian noise with SNR of (a) 0 dB and (b) 10 dB. The signal squared absolute value at frequency of 2 kHz is compared with its estimated variance using 5-state MSTF-GARCH model and by using the decision-directed approach.

It shows that the MSTF-GARCH approach with 5 states yields a better estimate of the squared absolute value both under high and low SNR conditions, especially in low energy bins. Furthermore, the MSTF-GARCH approach enables a better tracking of rapid changes in the coefficients energy than the decision-directed approach.

The differences between Figure 4.3 and Figure 4.4 is that the former demonstrates the *prediction* of the coefficients' variances (*i.e.*, the conditional variance) in a noiseless environment while the latter shows their second-order moments' estimation in a noisy environment. The variance prediction has a small delay of tracking rapid changes and the update step yields a better estimate of the squared absolute value in high energy bins. However, when noisy observations are employed, low-energy bins may be under the noise level and thus the estimation may be less accurate (for both the MSTF-GARCH approach and the decision-directed approach).

Figures 4.3 and 4.4 demonstrate that the proposed MSTF-GARCH model, when compared to a single-regime model, or to the decision-directed approach, improves the variance prediction and the squared absolute value estimation of speech signals in the STFT domain. Still, one needs to derive a frequency-dependent model and to estimate the signal presence probability in each time-frequency bin of the noisy speech signal based on the proposed model, which is a subject for further research.

4.7 Conclusions

We have proposed a statistical model for nonstationary processes with time-varying volatility structure in the STFT domain. Exploiting the advantages of both the conditional heteroscedasticity structure of GARCH models and the time-varying characteristics of hidden Markov chains, we model the expansion coefficients as multivariate, complex GARCH process with Markov-switching regimes. The correlation between successive coefficients in the time-frequency domain is taken into consideration by using the GARCH formulation which specifies the conditional variance as a linear function of its past values and past squared innovations. The time-varying structure of the conditional variance is determined by a hidden Markov chain which allows a different GARCH formulation in each state.

We showed that an ML estimate of the model can be practically obtained from training signals (assuming that the number of states is known), and developed a recursive algorithm for estimating the signal and its conditional variance in the STFT domain from its noisy observations. The conditional variance is recursively estimated for any regime by iterating propagation and update steps, while the evaluation of the regime conditional probabilities is based on the recursive correlation of the process. Experimental results demonstrate the improved performance of the proposed recursive algorithm compared to using an estimator which assumes a stationary process, even when the number of assumed regimes is smaller than the true number. When the number of assumed regimes approaches the true one, the recursive estimator yields comparable restoration results to those achievable by using the true model parameters. The conditional variance of an MSTF-GARCH process, as well as the instantaneous SNR on each frame, change over time. It is demonstrated that the recursive estimation approach has relatively small performance degradation compared to the theoretical estimation limit in the MMSE sense. Performance evaluation with real speech signals demonstrates better variance estimation when using a multi-regime model, compared to using a single-regime model, and improved squared absolute value estimation in a noisy environment compared to using the decision-directed approach.

Several extensions of this work, which may be interesting for further research, include analysis of the algorithm sensitivity to the number of the assumed states, the parameters values and the training set; generalization of the multivariate complex Markov-switching GARCH model, such that the conditional covariance matrix is not necessarily diagonal and the correlation between distinct frequency-bins is also taken into account; and finally estimation of the signal presence probability in the time-frequency domain and modification of the recursive signal estimation algorithm under signal presence uncertainty.

4.A Application of Markov-Switching GARCH Model to Speech Enhancement in Subbands[2]

In this appendix, we introduce an application of the Markov-switching GARCH model for spectral speech enhancement. A GARCH model is utilized with Markov switching regimes, where the parameters are assumed to be frequency variant. The model parameters are evaluated in each frequency subband and a special state (regime) is defined for the case where speech coefficients are absent or below a threshold level. The problem of speech enhancement under speech presence uncertainty is addressed and it is shown a soft voice activity detector may be inherently incorporated within the algorithm. Experimental results demonstrate the potential of our proposed model to improve noise reduction while retaining weak components of the speech signal.

4.A.1 Introduction

Statistical modeling of speech signals in the short-time Fourier transform (STFT) domain is of much interest in many speech enhancement applications. The Gaussian model [33] enables to derive useful estimators for the speech expansion coefficients such as the minimum mean-square error (MMSE) of the short-term spectral amplitude (STSA), as well as MMSE of the log-spectral amplitude (LSA) [33, 34]. Recently, a generalized autoregressive conditional heteroscedasticity (GARCH) model has been introduced for statistically modeling speech signals in the STFT domain [24]. However, the proposed model assumes that the parameters are both time and frequency invariant and it also requires an independent detector for speech activity in the time-frequency domain. A Markov-switching time-frequency GARCH (MSTF-GARCH) model has been proposed in [127] for modeling nonstationary signals in the time-frequency domain. Accordingly, the parameters are allowed to change in time according to the state of a hidden Markov chain (*e.g.*, switching between speech phonemes), but the parameters are still frequency-invariant. The model is estimated using training signals based on maximum likelihood (ML) approach and a recursive algorithm has been derived for conditional variance estimation and signal re-

[2]This appendix is based on [133].

construction from noisy observations. However, not only that different phonemes may result in different GARCH parameters, speech signals are generally characterized by different both volatility and energy levels in various frequency bands. Therefore, different parameters may better represent different frequency subbands.

In this appendix, we modify the MSTF-GARCH model by assuming different Markov chains in distinct subbands with identical state transition probabilities. The GARCH parameters are state dependent and frequency variant. We define an additional state for the case where speech coefficients are absent (or below a certain threshold level) and introduce parameter estimation method which is computationally more efficient than the traditional ML approach. Furthermore, the probability of the speech absence state can be used as a soft voice activity detector which is naturally generated in the reconstruction algorithm. Experimental results demonstrate improved noise reduction performance while preserving weak components of the speech signal.

Section 4.A.2 introduces the statistical model. In Section 4.A.3, we show how the model parameters can be estimated and in Section 4.A.4, we derive the speech enhancement algorithm based on the proposed model. Finally, in Section 4.A.5 we evaluate the performance of the proposed algorithm.

4.A.2 Model formulation

Let $\{X_{tk} \,|\, t = 0, 1, ...T-1, k = 0, 1, ..., K-1\}$ denote the coefficients of a speech signal in a STFT domain, where t is the time frame index and k is the frequency-bin index. Let $\{v_{tk}\}$ be iid complex Gaussian random variables with zero-mean and unit variance, let κ_n denote the nth frequency subband with $n \in \{1, 2, ..., N\}$ and $N < K$. An $(m+1)$-state hidden Markov chain is assumed for each frequency subband, denoted by $S_t(\kappa_n)$, with a realization $s_t(\kappa_n) \in \{0, 1, ..., m\}$ and state transition probabilities which are independent of the subband index. Let $\mathcal{I}^t$ denote all available information up to time t, *i.e.*, $\{X_{\tau k} \,|\, \tau = 0, 1, ..., t,\ k = 0, 1, ..., K-1\}$ and the regimes (states) path. Given the active state $S_t(\kappa_n) = s_t(\kappa_n)$, the *one-frame-ahead conditional variance* of the spectral coefficient X_{tk}, $k \in \kappa_n$ is defined by $\lambda_{tk|t-1,s_t} \triangleq E\left\{|X_{tk}|^2 \,|\, \mathcal{I}^{t-1}, s_t\right\}$, with $s_t = s_t(\kappa_n)$. The speech

spectral coefficients are assumed to follow an MSTF-GARCH process of order $(1,1)$ [127]:

$$X_{tk} = \sqrt{\lambda_{tk|t-1,s_t}}v_{tk}\,, \quad k \in \kappa_n \tag{4.62}$$

$$\lambda_{tk|t-1,s_t} = \lambda_{\min,n,s_t} + \alpha_{n,s_t}\left|X_{t-1,k}\right|^2 + \beta_{n,s_t}\left(\lambda_{t-1,k|t-2,s_{t-1}} - \lambda_{\min,n,s_{t-1}}\right)\,, \tag{4.63}$$

where $\lambda_{\min,n,s_t} > 0$ and $\alpha_{n,s_t}, \beta_{n,s_t} \geq 0$ are sufficient constrains for the positivity of the one-frame-ahead conditional variance, given that the initial conditions satisfy $\lambda_{0k|-1,s_0} \geq \lambda_{\min,n,s_0}$ for all $k \in \kappa_n$ and $s_0 = 0,1,...,m$. Note that the model formulation in [127] is slightly different. We assume that the parameters are frequency dependent while each $\lambda_{\min,n,s_t}$ defines the minimum value of the conditional variance in subband κ_n under $S_t(\kappa_n) = s_t$. Let $a_{s_{t-1},s_t} \triangleq p(S_t = s_t \,|\, S_{t-1} = s_{t-1})$, let π_s denotes the stationary probability of state s and let Ψ be an $(m+1)\times(m+1)$ matrix with elements

$$\psi_{s+1,\tilde{s}+1} = \frac{\pi_{\tilde{s}}}{\pi_s}a_{\tilde{s},s}\left(\alpha_{n,s} + \beta_{n,s}\right)\,, \quad s,\tilde{s} = 0,1,...,m\,. \tag{4.64}$$

Then, a necessary and sufficient condition for asymptotic wide-sense stationarity of the model defined in (4.62) and (4.63) is $\rho(\Psi) < 1$, where $\rho(\cdot)$ denotes spectral radius [128]. This condition is also necessary to ensure a finite second order moment for the process.

The unconditional expectation of the state-dependent one-frame-ahead conditional variance follows

$$\begin{aligned} E\left\{\lambda_{tk|t-1,s_t}\right\} &= \lambda_{\min,n,s_t} + \alpha_{n,s_t}E\left\{\left|X_{t-1,k}\right|^2 \mid s_t\right\} \\ &\quad + \beta_{n,s_t}E\left\{\lambda_{t-1,k|t-2,S_{t-1}} \mid s_t\right\} - \beta_{n,s_t}E\left\{\lambda_{\min,n,S_{t-1}} \mid s_t\right\} \end{aligned} \tag{4.65}$$

with

$$\begin{aligned} E\left\{\lambda_{\min,n,S_{t-1}} \mid s_t\right\} &= \sum_{s_{t-1}} p\left(s_{t-1} \mid s_t\right)\lambda_{\min,n,s_{t-1}} \\ &= \sum_{s_{t-1}} \frac{\pi_{s_{t-1}}}{\pi_{s_t}}a_{s_{t-1},s_t}\lambda_{\min,n,s_{t-1}}\,. \end{aligned} \tag{4.66}$$

Therefore, the stationary variance of the process is given by (see [128])

$$\lim_{t\to\infty} E\left\{\left|X_{tk}\right|^2\right\} = \pi\left(I_{m+1} - \Psi\right)^{-1}\tilde{\lambda}_{\min,n}\,, \tag{4.67}$$

where π is a row vector of the stationary probabilities, I_{m+1} is the identity matrix of order $m+1$,

$$\tilde{\lambda}_{\min,n} \triangleq \left[\tilde{\lambda}_{\min,n,0}, \tilde{\lambda}_{\min,n,1}, ..., \tilde{\lambda}_{\min,n,m}\right]^T \tag{4.68}$$

and

$$\tilde{\lambda}_{\min,n,s} \triangleq \lambda_{\min,n,s} - \frac{\beta_{n,s}}{\pi_s} \sum_{\tilde{s}} \pi_{\tilde{s}} a_{\tilde{s},s} \lambda_{\min,n,\tilde{s}} \,. \tag{4.69}$$

4.A.3 Model estimation

The estimation of a GARCH model with Markov regimes is generally obtained from a training set using ML approach [5, 13]. However, the maximization of the likelihood function is numerically unstable for multi-regime processes and only a local maxima can be generally obtained. Assuming an $(m+1)$-state Markov chain with GARCH of order $(1,1)$ in each regime, the maximization process generates $(m+1)^2$ variables for the transition probabilities and additional $3 \times (m+1)$ variables for the GARCH parameters in each regime. Speech signals in the STFT domain demonstrate different levels of magnitudes in different subbands and the coefficients are generally sparse. Therefore, we limit the conditional variances in each subband within a dynamic range of η_g dB and define a special state for speech absence hypothesis. Let $\zeta_g \triangleq \max_{t,k} |X_{tk}|^2$ and $\zeta_n \triangleq \max_{t,k\in\kappa_n} |X_{tk}|^2$ denote the global maximum energy and the local maximum energy of the coefficients (in subband κ_n), respectively. Then, for the speech absence state (namely, $s_t = 0$), we set

$$\lambda_{\min,n,0} = 10^{\log_{10}\zeta_g - \eta_g/10} \,,\; \alpha_{n,0} = \beta_{n,0} = 0 \,. \tag{4.70}$$

Under speech presence, a local dynamic range of η_ℓ dB ($\eta_\ell < \eta_g$) is assumed for the conditional variances. Furthermore, the parameters $\lambda_{\min,n,s}$, $s > 0$ are chosen to enable tracking any transients between different levels of magnitudes results in switching the active state. Without loss of generality, we sort the states according to the minimum variance level such that

$$\lambda_{\min,n,1} = \max\left\{\lambda_{\min,n,0}, 10^{\log_{10}\zeta_n - \eta_\ell/10}\right\} \,, \tag{4.71}$$

and for $s = 2, ..., m$, $\lambda_{\min,n,s}$ are log-spaced between $\lambda_{\min,n,1}$ and ζ_n. Each state practically represents different floor level for the spectral coefficients' variance. The parameters $\alpha_{n,s}, \beta_{n,s}$ for $s > 0$ set the volatility level of the conditional variance and they are chosen as follows. Assuming an immutable state s, the stationary variance follows

$$\lambda_{\infty,n,s} \triangleq \lim_{t\to\infty,\, k\in\kappa_n} \lambda_{tk|t-1,s} = \lambda_{\min,n,s} \frac{1-\beta_{n,s}}{1-\alpha_{n,s}-\beta_{n,s}} \tag{4.72}$$

provided that $\alpha_{n,s}+\beta_{n,s}<1$. Since different states are related to different dynamic ranges in ascending order, we constrain $\lambda_{\infty,n,s} \leq \lambda_{\min,n,s+1}$ and therefore

$$\frac{1-\beta_{n,s}}{1-\alpha_{n,s}-\beta_{n,s}} \leq \frac{\lambda_{\min,n,s+1}}{\lambda_{\min,n,s}}\,. \tag{4.73}$$

The autoregressive parameters, $\beta_{n,s}$, are chosen experimentally while the moving average parameters, $\alpha_{n,s}$, are chosen to satisfy equality in (4.73). Although the clean signal is assumed to be available for the model estimation, it is only the high energy values that are needed in each subband. These values can be practically estimated from the noisy coefficients using the spectral subtraction approach. The state transition probabilities can be estimated from test signals such that each active state is determined by the energy level of the subband.

4.A.4 Spectral enhancement of noisy speech

Let D_{tk} denote the spectral coefficients of a noise signal which is uncorrelated with the speech signal and assume that $D_{tk} \sim \mathcal{CN}\left(0,\sigma_{tk}^2\right)$. Let $Y_{tk} = X_{tk} + D_{tk}$ be the noisy observations and let $\mathcal{Y}^t \triangleq \{Y_{\tau k} \,|\, \tau = 0,1,...,t\,, k = 0,1,...,K-1\}$ denote the set of the observed coefficients up to time t. The noise variance σ_{tk}^2 is assumed to be known and it can be practically estimated using the improved minima controlled recursive averaging approach [71]. Reconstruction of the one-frame-ahead conditional variances of the speech coefficients is carried out recursively for each state by

$$\begin{aligned}\hat{\lambda}_{tk|t-1,s_t} &= \lambda_{\min,n,s_t} + \alpha_{n,s_t} E\left\{|X_{t-1,k}|^2 \,|\, \mathcal{Y}^{t-1}, s_t\right\} \\ &\quad + \beta_{n,s_t} E\left\{\lambda_{t-1,k|t-2,S_{t-1}} \,|\, \mathcal{Y}^{t-1}, s_t\right\} - \beta_{n,s_t} E\left\{\lambda_{\min,n,S_{t-1}} \,|\, \mathcal{Y}^{t-1}, s_t\right\}\end{aligned} \tag{4.74}$$

where

$$\begin{aligned}E\left\{|X_{t-1,k}|^2 \,|\, \mathcal{Y}^{t-1}, s_t\right\} &= \sum_{s_{t-1}} p\left(s_{t-1} \,|\, s_t, \mathcal{Y}^{t-1}\right) E\left\{|X_{t-1,k}|^2 \,|\, \mathcal{Y}^{t-1}, s_{t-1}\right\} \\ &\triangleq \sum_{s_{t-1}} p\left(s_{t-1} \,|\, s_t, \mathcal{Y}^{t-1}\right) \hat{\lambda}_{t-1,k|t-1,s_{t-1}}\,,\end{aligned} \tag{4.75}$$

$$E\left\{\lambda_{t-1,k|t-2,S_{t-1}} \,|\, \mathcal{Y}^{t-1}, s_t\right\} \simeq \sum_{s_{t-1}} p\left(s_{t-1} \,|\, s_t, \mathcal{Y}^{t-1}\right) \hat{\lambda}_{t-1,k|t-2,s_{t-1}} \tag{4.76}$$

and

$$E\left\{\lambda_{\min,n,S_{t-1}} \,|\, \mathcal{Y}^{t-1}, s_t\right\} = \sum_{s_{t-1}} p\left(s_{t-1} \,|\, s_t, \mathcal{Y}^{t-1}\right) \lambda_{\min,n,s_{t-1}}\,. \tag{4.77}$$

A detailed algorithm for the conditional variance restoration is described in [127].

Having an estimate for the speech coefficient's second order moment under each state, $\hat{\lambda}_{tk|t,s_t}$, estimates of the speech coefficients are obtained by minimizing the mean-square error of the log-spectral amplitude (LSA). Let

$$\hat{\xi}_{tk,s_t} \triangleq \frac{\hat{\lambda}_{tk|t,s_t}}{\sigma_{tk}^2}, \qquad \hat{\vartheta}_{tk,s_t} \triangleq \frac{\hat{\xi}_{tk,s_t}}{1+\hat{\xi}_{tk,s_t}} \cdot \frac{|Y_{tk}|^2}{\sigma_{tk}^2}. \tag{4.78}$$

Then, the LSA estimation of the speech coefficients is given by

$$\hat{X}_{tk} = Y_{tk} \prod_{s_t} G\left(\hat{\xi}_{tk,s_t}, \hat{\vartheta}_{tk,s_t}\right)^{p\left(s_t \,|\, \mathcal{Y}^t\right)}, \tag{4.79}$$

where

$$G\left(\xi, \vartheta\right) = \frac{\xi}{1+\xi} \exp\left(\frac{1}{2}\int_{\vartheta}^{\infty} \frac{e^{-t}}{t} dt\right) \tag{4.80}$$

is the LSA gain function [34] and the state probabilities, $p\left(s_t \,|\, \mathcal{Y}^t\right)$, are evaluated according to [127].

4.A.5 Experimental results and discussion

In this section, we demonstrate the application of the proposed model to speech enhancement and to speech presence probability estimation.

The enhancement evaluation includes two objective quality measures; segmental SNR and log-spectral distortion (LSD). The speech signals used in our evaluation are taken from the TIMIT database. The signals are sampled in 16 kHz, degraded by a nonstationary factory noise and transformed into the STFT domain using half overlapping Hamming windows of 32 msec length. Twenty subbands are considered with global and local dynamic ranges of $\eta_g = 50$ dB and $\eta_\ell = 20$ dB, and four-state Markov chains (*i.e.*, $m = 3$) for each subband. The autoregressive parameters used in our simulations are $\beta_{n,s} = 0.8$ for all n and $s > 0$. In each subband, the state persistence probability is 0.8 and $a_{s,\tilde{s}}$ are equally chosen for all $s \neq \tilde{s}$. Figure 1 demonstrates the spectrograms and waveforms of a clean signal, noisy signal with SNR of 5 dB, and the enhanced signal obtained by the proposed algorithm. It shows that the background noise is highly attenuated while weak speech components are retained, even while noise transients occur. Furthermore, the segmental SNR and the LSD are improved. A subjective study of speech spectrograms and

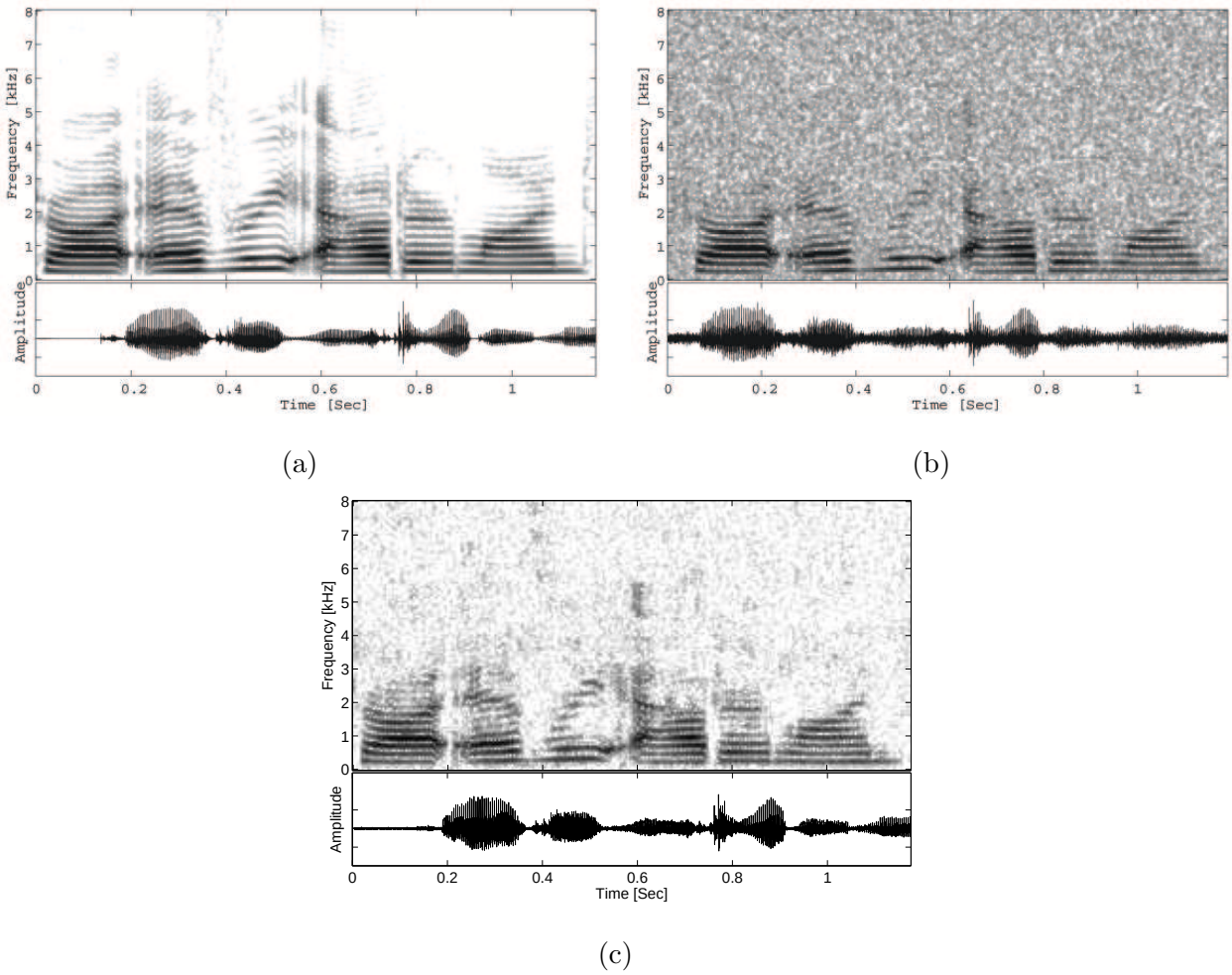


Figure 4.5: Speech spectograms and waveforms. (a) Clean signal: "Draw every outer line."; (b) speech corrupted by factory noise with 5 dB SNR (LSD= 6.68 dB, SegSNR= 0.05 dB); (c) speech reconstructed by using 4-state model (LSD= 3.14 dB, SegSNR= 6.76 dB).

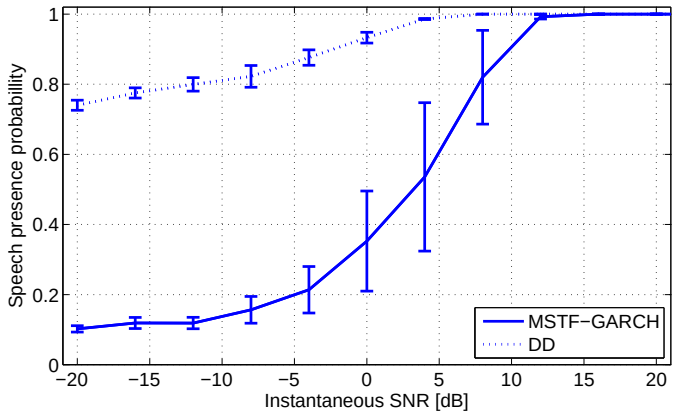


Figure 4.6: Conditional speech presence probability obtained by the proposed algorithm and by the decision-directed based algorithm.

informal listening tests confirm that the quality of the enhanced speech is improved by using frequency-dependent parameters which are derived from the different energy levels.

The conditional speech presence probability results from the enhancement algorithm is compared with the statistical model-based voice activity detector (with *hang-over*) of Sohn *et al.* [52] when applied to subbands. The later evaluates the conditional likelihood $\mathcal{L}_t \triangleq p\left(\mathcal{Y}^t \,|\, S_t \neq 0\right) / p\left(\mathcal{Y}^t \,|\, S_t = 0\right)$ by utilizing the decision-directed approach for the a priori SNR estimation (assuming only two states). The conditional speech presence probability is obtained by $p\left(S_t \neq 0 \,|\, \mathcal{Y}^t\right) = \mu\mathcal{L}_t / \left(1 + \mu\mathcal{L}_t\right)$, where $\mu \triangleq p\left(S_t \neq 0\right) / p\left(S_t = 0\right)$ is the *a priori* probabilities ratio. Figure 2 demonstrates the speech presence probabilities achieved when both algorithms are applied to a speech signal corrupted by a white Gaussian noise with SNR of 15 dB. The *instantaneous SNR* is defined as the ratio between the norms of the clean signal and the noise signal in each subband. It can be seen that the speech presence probability, derived from our proposed algorithm, results in a higher dynamic range for the probabilities and in much lower values for low energy coefficients. Furthermore, the probabilities ascribed to each instantaneous SNR are with higher variance resulting from the Markovian nature of the model.

Chapter 5

State Smoothing in Markov-Switching GARCH Models[1]

In this chapter, we address the problem of state smoothing in path-dependent Markov-switching generalized autoregressive conditional heteroscedasticity (GARCH) processes. We develop a smoothing algorithm which extends the *forward-backward recursions* of Chang and Hancock and the *stable backward recursion* of Lindgren, Askar and Derin. Two recursive steps are derived for the evaluation of conditional densities of future observations. The first step is an upward recursion which manipulates the future observations for the evaluation of their conditional densities, and the second step is a backward recursion which integrates over the possible future paths. Experimental results demonstrate the improvement in performance, compared to using causal estimation.

5.1 Introduction

State estimation is of both theoretical and practical importance whenever the underlying statistical model switches regimes over time [5, 129]. State smoothing (*i.e.*, noncausal state estimation) of hidden Markov processes (HMPs) has been originally introduced by Chang and Hancock [118]. Their solution for estimating the noncausal state probability, which is implemented using *forward-backward recursions*, decouples a forward recursion for the evaluation of the joint probability density of the current state and all observations

[1] This chapter is based on [130].

up to the same time, and a backward recursion for obtaining the future observations' density given the current state. Lindgren [119] and Askar and Derin [120] developed an alternative *stable backward recursion* for the state smoothing in HMPs. Kim [134] extended the stable backward recursion to nonmemoryless autoregressive hidden Markov processes (AR-HMPs) where both the current state (regime) and a finite set of past values are required for the conditional density evaluation, see also [5, chap. 22].

Generalized autoregressive conditional heteroscedasticity (GARCH) models and also Markov-switching GARCH (MS-GARCH) models, are widely used in the field of econometrics for volatility forecast derivation of economics rates [7, 8, 12, 13] and they have recently been utilized to several signal processing applications. In [135] GARCH modeling has been applied to spatially non-uniform noise in multichannel signal processing. In [26] a regime-switching GARCH model has been utilized for speech recognition and a complex-valued GARCH model has been proposed in [24, 25] for modeling speech signals in the short-time Fourier transform (STFT) domain for the application of speech enhancement. Generally, when incorporating GARCH processes with switching-regimes, the volatility evaluation requires knowledge of the pertinent history of the regime-switching GARCH process, including the regime-path [12, 13]. Properties of path-dependent MS-GARCH models have been studied by Francq et al. [122]. In order to estimate the model parameters, they showed that the conditional likelihood depends on all the possible paths and for a Markov-switching ARCH model (in which case there is no dependency on past active regimes) they showed that the forward-backward recursions can be employed for the conditional likelihood evaluation. The complex-valued GARCH model has been shown to be useful in speech enhancement applications [24, 25]. Motivated by extending the dynamic formulation of the time-frequency GARCH model and enabling a better fit for a process with a more complicated time-varying statistical behavior, a Markov-switching time-frequency GARCH (MSTF-GARCH) model has been introduced [127]. However, existing smoothing solutions are inapplicable in case of a path-dependent MS-GARCH model since both past observations and the regime path are required for the conditional variance estimation, whereas existing smoothing techniques rely on the assumption that given the current state, past active regimes are statistically independent of future densities.

In this chapter, we develop a state smoothing approach for MSTF-GARCH processes. The dependency of the conditional variance on past observations and past active regimes are taken into consideration as we generalize both the forward-backward recursions of Chang and Hancock [118] and the stable backward recursion of Lindgren [119] and Askar and Derin [120]. We derive two recursive steps for the evaluation of conditional densities of future observations. The first step is an upward recursion which manipulates the future observations for the evaluation of their conditional densities, corresponding to all possible future paths. The second step is a backward recursion which integrates over these paths to evaluate the future densities required for the noncausal state probability. The computational complexity of the generalized recursions grows exponentially with the number of future observations employed for the fixed-lag smoothing. However, experimental results demonstrate that the significant part of the improvement in performance, compared to using causal estimation, is achieved by considering a few future observations.

The organization of this chapter is as follows: In Section 5.2, we introduce the Markov-switching time-frequency GARCH model and formulate the state smoothing problem. In Section 5.3 we develop generalized forward-backward recursions as well as generalized stable backward recursion, and derive our noncausal state probability approach. Finally, in Section 5.4 we provide experimental results which demonstrate state smoothing for noisy Markov-switching time-frequency GARCH processes.

5.2 Problem formulation

Let $\mathbf{X}_t \in \mathbb{C}^K$ be a K-dimensional random vector at a discrete time t, and let X_{tk}, $k \in \{0, ..., K-1\}$ be its kth element. Let $\mathcal{X}_{t_1}^{t_2} = \{\mathbf{X}_t \,|\, t_1 \leq t \leq t_2\}$ represent the data set from time t_1 up to t_2 and let $\mathcal{X}^t \triangleq \mathcal{X}_0^t$. Let S_t denote the (unobserved) state at time t and let s_t be a realization of S_t, assuming S_t is a first-order Markov chain with transition probabilities $a_{s_t s_{t+1}} \triangleq p\,(S_{t+1} = s_{t+1} \,|\, S_t = s_t)$. Let $\mathcal{I}^t \triangleq \{\mathcal{X}^t, \mathcal{S}^t\}$ denote all available information up to time t, where $\mathcal{S}^t \triangleq \mathcal{S}_0^t = \{s_0, ..., s_t\}$. We assume that X_{tk} are generated by an m-state Markov-switching time-frequency GARCH process of order $(1, 1)$ which follows [127]:

$$X_{tk} = \sqrt{\lambda_{tk|t-1}} V_{tk}\,, \quad k = 0, ..., K-1\,, \tag{5.1}$$

where $\{V_{tk}\}$ are iid complex-valued random variables with zero-mean, unit variance and some known probability density. Given the state s_t, the conditional variance of X_{tk}, $\lambda_{tk|t-1,s_t} = E\{|X_{tk}|^2 \,|\, \mathcal{I}^{t-1}, s_t\}$, is a linear function of the previous conditional variance and squared absolute value:

$$\lambda_{tk|t-1} \equiv \lambda_{tk|t-1,s_t} = \xi_{s_t} + \alpha_{s_t} |X_{t-1,k}|^2 + \beta_{s_t} \lambda_{t-1,k|t-2}\,, \tag{5.2}$$

where $\xi_s > 0, \quad \alpha_s \geq 0$, and $\beta_s \geq 0, \quad s = 1, ..., m$ are sufficient constrains for the positivity of the conditional variance. Let Ψ be an m-by-m matrix with elements

$$\psi_{ij} = a_{ji}(\alpha_i + \beta_i)\pi_j/\pi_i \tag{5.3}$$

where $\pi_i = p(S_t = i)$ is the stationary probability of state i, and let $\rho(\cdot)$ represent the spectral radius of a matrix. Then, a necessary and sufficient condition for the process defined in (5.1) and (5.2) to be asymptotically wide-sense stationary is $\rho(\Psi) < 1$ [128].

Let $\mathbf{Y}_t = \mathbf{X}_t + \mathbf{D}_t$ denote the observed noisy signal, where $\mathbf{D}_t$ denotes the noise process which is uncorrelated with the signal $\mathbf{X}_t$, and let $\mathbf{D}_t$ be a zero-mean complex-valued Gaussian random process with a diagonal covariance matrix $E\left\{\mathbf{D}_t \mathbf{D}_t^H\right\} = \text{diag}\{\boldsymbol{\sigma}^2\}$, where $(\cdot)^H$ denotes the Hermitian transpose operation. The state conditional probability of a Markov-switching model, $p\left(s_t|\mathcal{Y}^\tau\right)$, is of considerable theoretical and practical importance for signal restoration and state sequence estimation (*e.g.,* [127, 129]).

Solutions of the state smoothing problem, *i.e.*, $\tau > t$, are normally obtained for HMPs using the forward-backward recursions [118] or the stable backward recursion [119, 120]. Extensions of these recursions for nonmemoryless AR-HMPs [134], [5, Chap. 22], are based on the quality that s_t and a finite set of past clean observations give complete statistical knowledge of future densities. However, in case of a path-dependent MS-GARCH model, a recursive formulation specifies the conditional distribution of the process as dependent on both past observations and the regime path, and therefore existing smoothing solutions are inapplicable.

5.3 State probability smoothing

In this Section, we develop the noncausal state probability for the model defined in (5.1) and (5.2). The smoothed probability is derived by generalizing both the forward-backward

recursions [118] and the stable backward recursion [119, 120].

5.3.1 Generalized forward-backward recursions

Assume that the conditional variance of the process is recursively estimated for any given state (*e.g.,* as proposed in [127]) and assume that the set of the recursively estimated conditional variances at time t, $\hat{\Lambda}_t \triangleq \left\{\hat{\boldsymbol{\lambda}}_{t|t-1,S_t} \,|\, S_t = 1, ..., m\right\}$, with the observed signal $\mathbf{Y}_t$ are sufficient statistics for the next conditional variance estimation for any given regime [24, 127]. Let $\hat{\boldsymbol{\lambda}}_{\tau_2|\tau_1,\mathcal{S}_{\tau_0}^{\tau_2}} = E\left\{\mathbf{X}_{\tau_2} \odot \mathbf{X}_{\tau_2}^* \,|\, \mathcal{S}_{\tau_0}^{\tau_2}, \mathcal{Y}^{\tau_1}\right\}$, $\tau_2 \geq \tau_1 > \tau_0$ denote the vector of estimated conditional variances at time τ_2 based on the observations up to time τ_1 and on the given set of active regimes $\mathcal{S}_{\tau_0}^{\tau_2}$, where $\odot$ denotes a term-by-term multiplication and * denotes complex conjugation. Let

$$g\left(\boldsymbol{\lambda}_{t|t-1,s_t}, \mathbf{Y}_t\right) \triangleq E\left\{\mathbf{X}_t \odot \mathbf{X}_t^* \,|\, S_t = s_t, \boldsymbol{\lambda}_{t|t-1,s_t}, \boldsymbol{Y}_t\right\} \tag{5.4}$$

where the function $g(\cdot)$ is determined based on the statistical model of $\{V_{tk}\}$ [24]. Define the generalized forward density by

$$\alpha\left(s_t, \mathcal{Y}^t\right) \triangleq f\left(s_t, \hat{\Lambda}_t, \mathbf{Y}_t\right) \tag{5.5}$$

and the generalized backward density by

$$\beta\left(\mathcal{Y}_{t+l}^{t+L} \,|\, \mathcal{S}_t^{t+l-1}, \mathcal{Y}^{t+l-1}\right) \triangleq f\left(\mathcal{Y}_{t+l}^{t+L} \,|\, \mathcal{S}_t^{t+l-1}, \hat{\Lambda}_t, \mathcal{Y}_t^{t+l-1}\right) . \tag{5.6}$$

Then, by substituting $l = 1$ we have

$$f\left(s_t, \mathcal{Y}_t^{t+L} \,|\, \hat{\Lambda}_t\right) = \alpha\left(s_t, \mathcal{Y}^t\right) \beta\left(\mathcal{Y}_{t+1}^{t+L} \,|\, s_t, \mathcal{Y}^t\right) , \tag{5.7}$$

and the noncausal state probability can be obtained by

$$\begin{aligned} p\left(s_t|\mathcal{Y}^{t+L}\right) &= p\left(s_t \,|\, \hat{\Lambda}_t, \mathcal{Y}_t^{t+L}\right) \\ &= \frac{\alpha\left(s_t, \mathcal{Y}^t\right) \beta\left(\mathcal{Y}_{t+1}^{t+L} \,|\, s_t, \mathcal{Y}^t\right)}{\sum_{s_t} \alpha\left(s_t, \mathcal{Y}^t\right) \beta\left(\mathcal{Y}_{t+1}^{t+L} \,|\, s_t, \mathcal{Y}^t\right)} . \end{aligned} \tag{5.8}$$

Proposition 5.1. *The generalized forward density of an MSTF-GARCH(1,1) process, $\alpha\left(s_t, \mathcal{Y}^t\right)$, satisfies the following recursion:*

$$\alpha\left(s_t, \mathcal{Y}^t\right) = f\left(\mathbf{Y}_t \,|\, s_t, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}\right) \sum_{s_{t-1}} \alpha\left(s_{t-1}, \mathcal{Y}^{t-1}\right) a_{s_{t-1}s_t} , \tag{5.9}$$

with the initial condition $\alpha\left(s_0, \mathbf{Y}_0\right) = p(s_0) f\left(\mathbf{Y}_0 \,|\, s_0\right)$.

Proof. The generalized forward density is obtained by

$$\alpha\left(s_t, \mathcal{Y}^t\right) = f\left(\mathbf{Y}_t \,|\, s_t, \hat{\Lambda}_t\right) f\left(s_t, \hat{\Lambda}_t\right) . \tag{5.10}$$

Given the active regime, the state-dependent conditional variance is sufficient for the conditional density. Furthermore, $\hat{\Lambda}_t$ and $\left\{\hat{\Lambda}_{t-1}, \mathbf{Y}_{t-1}\right\}$ represent the same statistical information. Hence

$$\alpha\left(s_t, \mathcal{Y}^t\right) = f\left(\mathbf{Y}_t \,|\, s_t, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}\right) f\left(s_t, \hat{\Lambda}_{t-1}, \mathbf{Y}_{t-1}\right) , \tag{5.11}$$

where

$$f\left(s_t, \hat{\Lambda}_{t-1}, \mathbf{Y}_{t-1}\right) = \sum_{s_{t-1}} \alpha\left(s_{t-1}, \mathcal{Y}^{t-1}\right) a_{s_{t-1}s_t} . \tag{5.12}$$

Substituting (5.12) into (5.11) we obtain the recursive formulation for the generalized forward density[2]. □

Proposition 5.2. *The generalized backward density of an MSTF-GARCH(1,1) process,* $\beta\left(\mathcal{Y}_{t+1}^{t+L} \,|\, s_t, \mathcal{Y}^t\right)$*, satisfies the following two-step recursion:*

Step I: For $l = 1, ..., L$ *and all* $\mathcal{S}_t^{t+l}$*:*

$$\hat{\boldsymbol{\lambda}}_{t+l|t+l-1,\mathcal{S}_t^{t+l}} = \xi_{s_{t+l}}\mathbf{1} + \alpha_{s_{t+l}}\hat{\boldsymbol{\lambda}}_{t+l-1|t+l-1,\mathcal{S}_t^{t+l-1}} + \beta_{s_{t+1}}\hat{\boldsymbol{\lambda}}_{t+l-1|t+l-2,\mathcal{S}_t^{t+l-1}} \tag{5.13}$$

$$\hat{\boldsymbol{\lambda}}_{t+l|t+l,\mathcal{S}_t^{t+l}} = g\left(\hat{\boldsymbol{\lambda}}_{t+l|t+l-1,\mathcal{S}_t^{t+l}}, \mathbf{Y}_{t+l}\right) . \tag{5.14}$$

Step II: For $l = L, ..., 1$ *and all* $\mathcal{S}_t^{t+l}$*:*

$$f\left(\mathcal{Y}_{t+l}^{t+L} \,|\, \mathcal{S}_t^{t+l}, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}, \mathcal{Y}_t^{t+l-1}\right) = \beta\left(\mathcal{Y}_{t+l+1}^{t+L} \,|\, \mathcal{S}_t^{t+l}, \mathcal{Y}^{t+l}\right) f\left(\mathbf{Y}_{t+l} \,|\, \mathcal{S}_t^{t+l}, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}, \mathcal{Y}_t^{t+l-1}\right) \tag{5.15}$$

$$\beta\left(\mathcal{Y}_{t+l}^{t+L} \,|\, \mathcal{S}_t^{t+l-1}, \mathcal{Y}^{t+l-1}\right) = \sum_{s_{t+l}} f\left(\mathcal{Y}_{t+l}^{t+L} \,|\, \mathcal{S}_t^{t+l}, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}, \mathcal{Y}_t^{t+l-1}\right) a_{s_{t+l-1}s_{t+l}} , \tag{5.16}$$

with $\beta\left(\mathcal{Y}_{t+L+1}^{t+L} \,|\, \mathcal{S}_t^{t+L}, \mathcal{Y}^{t+L}\right) = 1$ *as the initial condition for the second step, and where* $\mathbf{1}$ *denotes a vector of ones.*

[2]The initial conditions for the generalized forward recursion have negligible effect on the conditional densities assuming an asymptotic stationary process which is sufficiently long. Therefore, the initial conditional variance $f\left(\mathbf{Y}_0 \,|\, s_0\right)$ can be estimated by using the state-dependent stationary density of the process.

Proof. The generalized backward density $\beta\left(\mathcal{Y}_{t+1}^{t+L} \,|\, s_t, \mathcal{Y}^t\right) = f\left(\mathcal{Y}_{t+1}^{t+L} \,|\, s_t, \mathcal{Y}^t\right)$ can be obtained by

$$\beta\left(\mathcal{Y}_{t+1}^{t+L} \,|\, s_t, \mathcal{Y}^t\right) = \sum_{s_{t+1}} f\left(\mathcal{Y}_{t+1}^{t+L} \,|\, \mathcal{S}_t^{t+1}, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}, \mathbf{Y}_t\right) a_{s_t s_{t+1}}\,, \tag{5.17}$$

where the multivariate density $f\left(\mathcal{Y}_{t+1}^{t+L} \,|\, \mathcal{S}_t^{t+1}, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}, \mathbf{Y}_t\right)$ in (5.17) can be obtained by

$$f\left(\mathcal{Y}_{t+1}^{t+L} \,|\, \mathcal{S}_t^{t+1}, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}, \mathbf{Y}_t\right) = \beta\left(\mathcal{Y}_{t+2}^{t+L} \,|\, \mathcal{S}_t^{t+1}, \mathcal{Y}^{t+1}\right) f\left(\mathbf{Y}_{t+1} \,|\, \mathcal{S}_t^{t+1}, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}, \mathbf{Y}_t\right)\,. \tag{5.18}$$

From (5.17) and (5.18) we recursively obtain for any $l = 1, ..., L$:

$$\beta\left(\mathcal{Y}_{t+l}^{t+L} \,|\, \mathcal{S}_t^{t+l-1}, \mathcal{Y}^{t+l-1}\right) = \sum_{s_{t+l}} f\left(\mathcal{Y}_{t+l}^{t+L} \,|\, \mathcal{S}_t^{t+l}, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}, \mathcal{Y}_t^{t+l-1}\right) a_{s_{t+l-1} s_{t+l}} \tag{5.19}$$

and

$$f\left(\mathcal{Y}_{t+l}^{t+L} \,|\, \mathcal{S}_t^{t+l}, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}, \mathcal{Y}_t^{t+l-1}\right) = \beta\left(\mathcal{Y}_{t+l+1}^{t+L} \,|\, \mathcal{S}_t^{t+l}, \mathcal{Y}^{t+l}\right) f\left(\mathbf{Y}_{t+l} \,|\, \mathcal{S}_t^{t+l}, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}, \mathcal{Y}_t^{t+l-1}\right)\,. \tag{5.20}$$

The conditional density $f\left(\mathbf{Y}_{t+l} \,|\, \mathcal{S}_t^{t+l}, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}, \mathcal{Y}_t^{t+l-1}\right)$ in (5.20) is the density of the observed data at time $t+l$ conditioned on the regime path $\mathcal{S}_t^{t+l}$, the recursively estimated conditional variance at time t given s_t, and also on all observations from time t up to time $t+l-1$. This density has a diagonal covariance matrix with the following conditional variance on its diagonal:

$$\begin{aligned} E\left\{\mathbf{Y}_{t+l} \odot \mathbf{Y}_{t+l}^* \,|\, \mathcal{S}_t^{t+l}, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}, \mathcal{Y}_t^{t+l-1}\right\} & \\ = \boldsymbol{\sigma}^2 + \hat{\boldsymbol{\lambda}}_{t+l|t+l-1,\mathcal{S}_t^{t+l}} & \\ = \boldsymbol{\sigma}^2 + \xi_{s_{t+l}}\mathbf{1} + \alpha_{s_{t+l}} E\{\mathbf{X}_{t+l-1} \odot \mathbf{X}_{t+l-1}^* \,|\, \mathcal{S}_t^{t+l}, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}, \mathcal{Y}_t^{t+l-1}\} & \\ + \beta_{s_{t+l}} E\{\boldsymbol{\lambda}_{t+l-1|t+l-2} \,|\, \mathcal{S}_t^{t+l-1}, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}, \mathcal{Y}_t^{t+l-2}\}. & \end{aligned} \tag{5.21}$$

The expected absolute squared value of the signal at a specific time given the active regime is independent of any future regimes, hence

$$\begin{aligned} E\left\{\mathbf{X}_{t+l-1} \odot \mathbf{X}_{t+l-1}^* \,|\, \mathcal{S}_t^{t+l}, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}, \mathcal{Y}_t^{t+l-1}\right\} &= \hat{\boldsymbol{\lambda}}_{t+l-1|t+l-1,\mathcal{S}_t^{t+l-1}} \\ &= g\left(\hat{\boldsymbol{\lambda}}_{t+l-1|t+l-2,\mathcal{S}_t^{t+l-1}}, \mathbf{Y}_{t+l-1}\right) \end{aligned} \tag{5.22}$$

Combining (5.22) with (5.21), we obtain Step I of the generalized backward recursion ((5.13) and (5.14)), and from (5.19) and (5.20) we obtain Step II ((5.15) and (5.16)). □

Step I is an upward recursion which manipulates the future observations for estimating their conditional variances corresponding to all possible regime sequences $\mathcal{S}_t^{t+L}$. Step II is a backward recursion, which integrates the Step I results to evaluate the generalized backward density. Each step of the generalized backward recursion is calculated for m^{L+1} regime sequences, and therefore the computational complexity increases exponentially with the delay L. However, as the correlation of the current state and future observations decreases along time, small values of L sufficiently enhance the chain sequence estimation, as can be seen in the experimental results.

5.3.2 Generalized stable backward recursion

The stable backward recursion is derived by using the smoothed probability of two sequential states, which is given by [120]:

$$p\left(\mathcal{S}_t^{t+1} \,|\, \mathcal{Y}^{t+L}\right) = \frac{f\left(\mathcal{S}_t^{t+1}, \mathcal{Y}_{t+1}^{t+L} \,|\, \mathcal{Y}^{t}\right) f\left(s_{t+1} \,|\, \mathcal{Y}^{t+L}\right)}{f\left(s_{t+1}, \mathcal{Y}_{t+1}^{t+L} \,|\, \mathcal{Y}^{t}\right)}. \tag{5.23}$$

Under the assumption that $\left\{\hat{\Lambda}_t, \mathbf{Y}_t\right\}$ are sufficient statistics for the next state-dependent conditional variance estimation, we obtain

$$\begin{aligned} f\left(\mathcal{S}_t^{t+1}, \mathcal{Y}_{t+1}^{t+L} \,|\, \mathcal{Y}^{t}\right) &= f\left(s_{t+1}, \mathcal{Y}_{t+1}^{t+L} \,|\, s_t, \mathcal{Y}^{t}\right) p\left(s_t \,|\, \mathcal{Y}^{t}\right) \\ &= f\left(\mathcal{Y}_{t+1}^{t+L} \,|\, \mathcal{S}_t^{t+1}, \mathcal{Y}^{t}\right) p\left(s_{t+1} \,|\, s_t, \mathcal{Y}^{t}\right) p\left(s_t \,|\, \mathcal{Y}^{t}\right) \\ &= f\Big(\mathcal{Y}_{t+1}^{t+L} \,|\, \mathcal{S}_t^{t+1}, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}, \mathbf{Y}_t\Big)\, a_{s_t s_{t+1}} p\Big(s_t \,|\, \hat{\Lambda}_t, \mathbf{Y}_t\Big) \end{aligned} \tag{5.24}$$

and

$$\begin{aligned} f\left(s_{t+1}, \mathcal{Y}_{t+1}^{t+L} \,|\, \mathcal{Y}^{t}\right) &= f\left(\mathcal{Y}_{t+1}^{t+L} \,|\, s_{t+1}, \mathcal{Y}^{t}\right) p\left(s_{t+1} \,|\, \mathcal{Y}^{t}\right) \\ &= f\left(\mathcal{Y}_{t+1}^{t+L} \,|\, s_{t+1}, \hat{\boldsymbol{\lambda}}_{t+1|t,s_{t+1}}\right) p\left(s_{t+1} \,|\, \hat{\Lambda}_t, \mathbf{Y}_t\right). \end{aligned} \tag{5.25}$$

By substituting (5.24) and (5.25) into (5.23) and integrating out all states at time $t+1$, we obtain the following backward recursion for the smoothed state probability:

$$p\left(s_t \,|\, \mathcal{Y}^{t+L}\right) = p\left(s_t \,|\, \hat{\Lambda}_t, \mathbf{Y}_t\right) \sum_{s_{t+1}} \frac{f\left(\mathcal{Y}_{t+1}^{t+L} \,|\, \mathcal{S}_t^{t+1}, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}, \mathbf{Y}_t\right) a_{s_t s_{t+1}} p\left(s_{t+1} \,|\, \mathcal{Y}^{t+L}\right)}{f\left(\mathcal{Y}_{t+1}^{t+L} \,|\, s_{t+1}, \hat{\boldsymbol{\lambda}}_{t+1|t,s_{t+1}}\right) p\left(s_{t+1} \,|\, \hat{\Lambda}_t, \mathbf{Y}_t\right)}, \tag{5.26}$$

where the conditional density $f\left(\mathcal{Y}_{t+1}^{t+L} \,|\, \mathcal{S}_t^{t+1}, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}, \mathbf{Y}_t\right)$ can be derived from the generalized backward recursion (5.13)-(5.16). However, the conditional density

$f\left(\mathcal{Y}_{t+1}^{t+L} \mid s_{t+1}, \hat{\boldsymbol{\lambda}}_{t+1|t,s_{t+1}}\right)$ in the denominator of (5.26) requires calculation of a similar recursion which is not informed of the regime s_t.

Although the stable backward recursion is known to be numerically more stable than the forward-backward recursions, the instability of the latter is insignificant for short delays and the former requires computation of the generalized backward recursion twice, one for evaluating $f\left(\mathcal{Y}_{t+1}^{t+L} \mid \mathcal{S}_t^{t+1}, \hat{\boldsymbol{\lambda}}_{t|t-1,s_t}, \mathbf{Y}_t\right)$ and one for $f\left(\mathcal{Y}_{t+1}^{t+L} \mid s_{t+1}, \hat{\boldsymbol{\lambda}}_{t+1|t,s_{t+1}}\right)$.

5.4 Experimental results

The generalized state smoothing has been applied to state detection in noisy MSTF-GARCH(1, 1) processes with 3 states and 5 to 15 dB signal-to-noise ratios (SNRs). Twenty random stationary models have been simulated with an unconditional Gaussian model and uniformly distributed parameters on the intervals $(0, 1/3]$, $(1/3, 2/3]$ and $(2/3, 1]$ for each state respectively. For each model 20 signals are considered, each of dimension $K = 100$ and time length $T = 100$. The conditional variances $\hat{\boldsymbol{\lambda}}_{t|t-1,s_t}$ are estimated using the recursive approach of [127]. Figure 5.1 shows the detection error rate $p\left(\hat{s}_t \neq s_t\right)$ for casual estimation as well as for noncausal estimation with up to $L = 4$ samples delay. It can be seen that the state detection monotonically improves with the increase of the delay. However, the most significant improvement is achieved by using up to 2 future samples, and the contribution of additional future observations decays along time.

5.5 Conclusions

We have derived state smoothing for Markov-switching time-frequency GARCH process, in which case the conditional variances depend on both past observations and the regime path. Our noncausal state probability solution generalizes both the standard forward-backward recursions and the stable backward recursion of HMP by capturing both the signal correlation along time and its conditioning on the regime path. Accordingly, the backward recursion requires two recursive steps for evaluating the conditional density of the given future observations corresponding to all optional future paths. Although the computational complexity of the generalized backward recursion grows exponentially

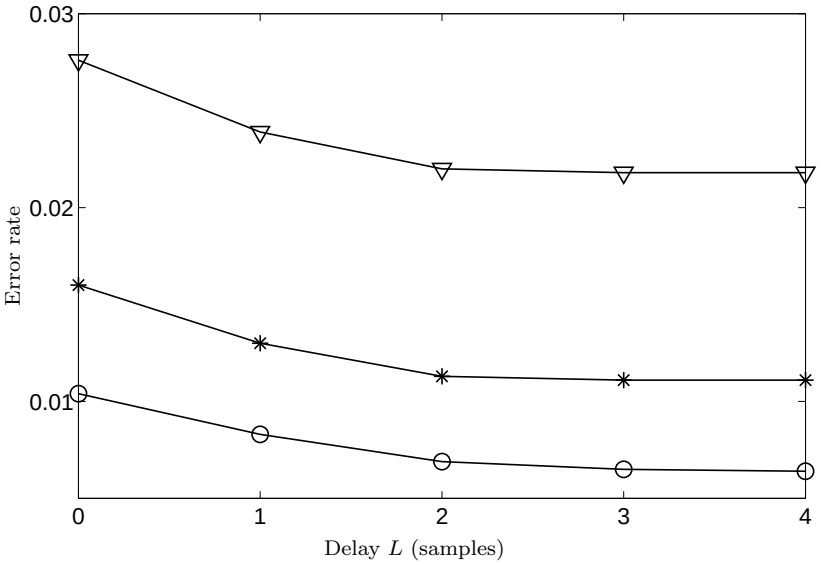


Figure 5.1: State smoothing error rate for 3-state MSTF-GARCH models with SNRs of 5 dB (triangle), 10 dB (asterisk) and 15 dB (circle).

with the delay, a small number of future observations contribute with the most significant improvement to the state estimation. Combining the generalized recursions with the recursive signal restoration algorithm of [127] facilitates a noncausal signal restoration, which is a subject for further research.

Chapter 6

Simultaneous Detection and Estimation Approach for Speech Enhancement[1]

In this chapter, we present a simultaneous detection and estimation approach for speech enhancement. A detector for speech presence in the short-time Fourier transform domain is combined with an estimator, which jointly minimizes a cost function that takes into account both detection and estimation errors. Cost parameters control the trade-off between speech distortion, caused by missed detection of speech components, and residual musical noise resulting from false-detection. Furthermore, a modified decision-directed *a priori* signal-to-noise ratio (SNR) estimation is proposed for transient-noise environments. Experimental results demonstrate the advantage of using the proposed simultaneous detection and estimation approach with the proposed *a priori* SNR estimator, which facilitate suppression of transient noise with a controlled level of speech distortion.

In Appendix 6.B we formulate a speech enhancement problem under multiple hypotheses, assuming an indicator or detector for the transient noise presence is available in the short-time Fourier transform (STFT) domain. Hypothetical presence of speech or transient noise is considered in the observed spectral coefficients, and cost parameters control the trade-off between speech distortion and residual transient noise. An optimal estimator, which minimizes the mean-square error of the log-spectral amplitude, is derived,

[1]This chapter is based on [136].

while taking into account the probability of erroneous detection. Experimental results demonstrate the improved performance in transient noise suppression, compared to using the optimally-modified log-spectral amplitude estimator.

6.1 Introduction

Optimal design of efficient speech enhancement algorithms has attracted significant research effort for several decades. Speech enhancement systems often operate in the short-time Fourier transform (STFT) domain, where the speech spectral coefficients are estimated from the spectral coefficients of the degraded signal. The spectral coefficients of the speech signal are generally sparse in the STFT domain in the sense that speech is present only in some of the frames, and in each frame only some of the frequency-bins contain the significant part of the signal energy. However, existing algorithms often focus on estimating the spectral coefficients rather than detecting their existence. The spectral-subtraction algorithm [29, 30] contains an elementary detector for speech activity in the time-frequency domain, but it generates musical noise caused by falsely detecting noise peaks as bins that contain speech, which are randomly scattered in the STFT domain. Subspace approaches for speech enhancement [57, 59, 60, 104] decompose the vector of the noisy signal into a signal-plus-noise subspace and a noise subspace, and the speech spectral coefficients are estimated after removing the noise subspace. Accordingly, these algorithms are aimed at detecting the speech coefficients and subsequently estimating their values. McAulay and Malpass [32] were the first to propose a speech spectral estimator under a two-state model. They derived a maximum likelihood (ML) estimator for the speech spectral amplitude under speech-presence uncertainty. Ephraim and Malah followed this approach of signal estimation under speech presence uncertainty and derived an estimator which minimizes the mean-square error (MSE) of the short-term spectral amplitude (STSA) [33]. In [49], speech presence probability is evaluated to improve the minimum MSE (MMSE) of the log-spectral amplitude (LSA) estimator, and in [38] a further improvement of the MMSE-LSA estimator is achieved based on a two-state model. Under speech absence hypothesis Cohen and Berdugo [38] considered a constant attenuation factor to enable a more natural residual noise, characterized by reduced musicality.

Under slowly time-varying noise conditions, an estimator which minimizes the MSE of the STSA or the LSA under speech presence uncertainty may yield reasonable results [33, 38]. However, under quickly time-varying noise conditions, abrupt transients may not be sufficiently attenuated, since speech is falsely detected with some positive probability. Reliable detectors for speech activity and noise transients are necessary to further attenuate noise transients without much degrading the speech components [107,137]. Despite the sparsity of speech coefficients in the time-frequency domain and the importance of signal detection for noise suppression performance, common speech enhancement algorithms deal with speech detection *independently* of speech estimation. Even when a voice activity detector is available in the STFT domain (*e.g.*, [51–56, 108]), it is not straightforward to consider the detection errors when designing the optimal speech estimator. High attenuation of speech spectral coefficients due to missed detection errors may significantly degrade speech quality and intelligibility, while falsely detecting noise transients as speech-contained bins, may produce annoying musical noise.

In this chapter, we present a novel formulation of the speech enhancement problem, which incorporates simultaneous operations of detection and estimation. A detector for the speech coefficients is combined with an estimator, which jointly minimizes a cost function that takes into account both estimation and detection errors. Under speech-presence, the cost is proportional to a quadratic spectral amplitude (QSA) error [33], while under speech-absence, the distortion depends on a certain attenuation factor [29, 38, 70]. We derive a combined detector and estimator with cost parameters that enable to control the trade-off between speech distortion, caused by missed detection of speech components, and residual musical noise resulting from false-detection. The combined solution generalizes the well-known STSA algorithm, which involves merely estimation under signal presence uncertainty. In addition, we propose a modification of the decision-directed *a priori* signal-to-noise ratio (SNR) estimator, which is suitable for transient-noise environments. Experimental results show that the simultaneous detection and estimation yields better noise reduction than the STSA algorithm while not degrading the speech signal. The advantage of using a suitable indicator for transient noise is demonstrated in a nonstationary noise environment, where the proposed algorithm facilitates suppression of transient noise with a controlled level of speech distortion.

The chapter is organized as follows. In Section 6.2, we briefly review classical speech enhancement under signal presence uncertainty. In Section 6.3, we reformulate the speech enhancement problem in the STFT domain as a simultaneous detection and estimation problem. In Section 6.4, we derive the combined solution for a QSA distortion function. In Section 6.5, we relate our proposed approach to the spectral-subtraction approach. In Section 6.6, we present an *a priori* SNR estimator suitable for transient noise environments, and in Section 6.7 we demonstrate the performance of the proposed approach compared to existing algorithms, both under stationary and transient-noise environments.

6.2 Classical speech enhancement

In this section, we present the classical approach for spectral speech enhancement in non-stationary noise environments, assuming that some indicator for transient noise activity is available.

Let $x(n)$ and $d(n)$ denote speech and uncorrelated additive noise signals, and let $y(n) = x(n) + d(n)$ be the observed signal. Applying the STFT to the observed signal, we have

$$Y_{\ell k} = X_{\ell k} + D_{\ell k}\,, \tag{6.1}$$

where $\ell = 0, 1, ...$ is the time frame index and $k = 0, 1, ..., K-1$ is the frequency-bin index. Let $H_1^{\ell k}$ and $H_0^{\ell k}$ denote, respectively, speech presence and absence hypotheses in the time-frequency bin (ℓ, k), *i.e.*,

$$\begin{aligned} H_1^{\ell k} &: Y_{\ell k} = X_{\ell k} + D_{\ell k} \\ H_0^{\ell k} &: Y_{\ell k} = D_{\ell k}\,. \end{aligned} \tag{6.2}$$

We assume that the noise expansion coefficients can be represented as the sum of two uncorrelated noise components $D_{\ell k} = D_{\ell k}^s + D_{\ell k}^t$, where $D_{\ell k}^s$ denotes a quasi-stationary noise component and $D_{\ell k}^t$ denotes a highly nonstationary transient component. The transient components are generally rare, but they may be of high energy and thus cause significant degradation to speech quality and intelligibility. However, in many applications, a reliable indicator for the transient noise activity may be available in the system. For example, in an emergency car (*e.g.*, police or ambulance) the engine noise may be

considered as quasi-stationary, but activating a siren results in a highly nonstationary noise which is perceptually very annoying. Since the sound generation in the siren is nonlinear, linear echo cancelers, *e.g.*, [138], may be inappropriate. In a computer-based communication system, a transient noise such as a keyboard typing noise may be present in addition to quasi-stationary background office noise. Another example is a digital camera, where activating the lens-motor (zooming in/out) may result in high-energy transient noise components, which degrade the recorded audio. In the above examples, an indicator for the transient noise activity may be available, *i.e.*, siren source signal, keyboard output signal and the lens-motor controller output. Furthermore, given that a transient noise source is active, a detector for the transient noise in the STFT domain may be designed and its spectrum can be estimated based on training data.

The objective of a speech enhancement system is to reconstruct the spectral coefficients of the speech signal such that under speech-presence a certain distortion measure between the spectral coefficient and its estimate, $d\left(X_{\ell k}, \hat{X}_{\ell k}\right)$, is minimized, and under speech-absence a constant attenuation of the noisy coefficient would be desired to maintain a natural background noise [38, 70]. Although the speech expansion coefficients are not necessarily present, most classical speech enhancement algorithms try to estimate the spectral coefficients rather than detecting their existence, or try to independently design detectors and estimators. The well-known spectral subtraction algorithm estimates the speech spectrum by subtracting the estimated noise spectrum from the noisy squared absolute coefficients [29, 30], and thresholding the result by some desired residual noise level. Thresholding the spectral coefficients is in fact a detection operation in the time-frequency domain, in the sense that speech coefficients are assumed to be absent in the low-energy time-frequency bins and present in noisy coefficients whose energy is above the threshold.

McAulay and Malpass were the first to propose a two-state model for the speech signal in the time-frequency domain [32]. Accordingly, the MMSE estimator follows [115]

$$\begin{aligned}\hat{X}_{\ell k} &= E\left\{X_{\ell k} \,|\, Y_{\ell k}\right\} \\ &= E\left\{X_{\ell k} \,|\, Y_{\ell k}, H_1^{\ell k}\right\} p\left(H_1^{\ell k} \,|\, Y_{\ell k}\right) \ . \end{aligned} \tag{6.3}$$

The resulting estimator does not detect speech components, but rather, a soft-decision

is performed to further attenuate the signal estimate by the *a posteriori* speech presence probability. Ephraim and Malah followed the same approach and derived an estimator which minimizes the MSE of the STSA under signal presence uncertainty [33]. Accordingly,

$$\left|\hat{X}_{\ell k}\right| = E\left\{|X_{\ell k}| \mid Y_{\ell k}, H_1^{\ell k}\right\} p\left(H_1^{\ell k} \mid Y_{\ell k}\right) . \tag{6.4}$$

Both in [32] and [33], under $H_0^{\ell k}$ the speech components are assumed zero and the *a priori* probability of speech presence is both time and frequency invariant, *i.e.*, $p\left(H_1^{\ell k}\right) = p\left(H_1\right)$. In [38, 49], the speech presence probability is evaluated for each frequency-bin and time-frame to improve the performance of the MMSE-LSA estimator [34]. Further improvement of the MMSE-LSA suppression rule can be achieved by considering under $H_0^{\ell k}$ a constant attenuation factor $G_f << 1$, which is determined by subjective criteria for residual noise naturalness, see also [70]. The OM-LSA estimator [38] is given by

$$\left|\hat{X}_{\ell k}\right| = \left(\exp\left[E\left\{\log|X_{\ell k}| \mid Y_{\ell k}, H_1^{\ell k}\right\}\right]\right)^{p\left(H_1^{\ell k} \mid Y_{\ell k}\right)} \left(G_f \ |Y_{\ell k}|\right)^{1-p\left(H_1^{\ell k} \mid Y_{\ell k}\right)} . \tag{6.5}$$

Suppose that an indicator for the presence of transient noise components is available in a highly nonstationary noise environment, then high-energy transients may be attenuated by using one of the above-mentioned estimators (6.3)–(6.5) and heuristically setting the *a priori* speech presence probability $p\left(H_1^{\ell k}\right)$ to a sufficiently small value. Unfortunately, this also results in suppression of desired speech components and intolerable degradation of speech quality. In general, an estimation-only approach under signal presence uncertainty produces larger speech degradation for small $p\left(H_1^{\ell k}\right)$, since the optimal estimate is attenuated by the *a posteriori* speech presence probability. On the other hand, increasing $p\left(H_1^{\ell k}\right)$ prevents the estimator from sufficiently attenuating noise components. Integrating a jointly optimal detector and estimator into the speech enhancement system may significantly improve the speech enhancement performance under highly non-stationary noise conditions and may allow further reduction of transient components without much degradation of the desired signal.

6.3 Reformulation of the speech enhancement problem

In this section, we reformulate the speech enhancement as a simultaneous detection and estimation problem.

Middleton and Esposito [115] were the first to propose simultaneous signal detection and estimation within the framework of statistical decision theory. A decision space, $\left\{\eta_0^{\ell k}, \eta_1^{\ell k}\right\}$, is assumed for the detection operation where under the decision $\eta_j^{\ell k}$, signal hypothesis $H_j^{\ell k}$ is accepted and a corresponding estimate $\hat{X}_{\ell k} = \hat{X}_{\ell k,j}$ is considered. The detection and estimation are strongly coupled so that the detector is optimized with the knowledge of the specific structure of the estimator, and the estimator is optimized in the sense of minimizing a Bayesian risk associated with the combined operations. For notation simplification, we omit the time-frequency indices (ℓ, k). Let

$$C_j\left(X, \hat{X}\right) \geq 0 \tag{6.6}$$

denote the cost of making a decision η_j (and choosing an estimator $\hat{X}_j$) where X is the desired signal. Then, the Bayes risk of the two operations associated with simultaneous detection and estimation is defined by [115, 116]

$$R = \sum_{j=0}^{1} \int_{\Omega_y} \int_{\Omega_x} C_j\left(X, \hat{X}\right) p\left(\eta_j \mid Y\right) p\left(Y \mid X\right) p\left(X\right) dX dY \tag{6.7}$$

where Ω_x and Ω_y are the spaces of the speech and noisy signals, respectively. The simultaneous detection and estimation approach is aimed at jointly minimizing the Bayes risk over both the decision rule and the corresponding signal estimate. Let $q \triangleq p\left(H_1\right)$ denote the *a priori* speech presence probability and let X_R and X_I denote the real and imaginary parts of the expansion coefficient X. Then, the *a priori* distribution of the speech expansion coefficient follows

$$p\left(X\right) = q\, p\left(X \mid H_1\right) + \left(1 - q\right) p\left(X \mid H_0\right) \,, \tag{6.8}$$

where $p\left(X \mid H_0\right) = \delta\left(X\right)$ and $\delta\left(X\right) \triangleq \delta\left(X_R, X_I\right)$ denotes the Dirac-delta function. The cost function $C_j\left(X, \hat{X}\right)$ may be defined differently whether H_1 or H_0 is true. Therefore,

we let

$$C_{ij}\left(X,\hat{X}\right) \triangleq C_j\left(X,\hat{X}\,|\,H_i\right) \tag{6.9}$$

denote the cost which is conditioned on the true hypothesis[2]. The cost function $C_{ij}\left(X,\hat{X}\right)$ depends on both the true signal value and its estimate under the decision η_j and therefore couples the operations of detection and estimation. By substituting (6.8) into (6.7) we obtain

$$\begin{aligned} R = \int_{\Omega_y}\int_{\Omega_x} p\left(Y\,|\,X\right)\Big\{ & p\left(\eta_0\,|\,Y\right)\Big[q\,p\left(X\,|\,H_1\right)C_{10}\left(X,\hat{X}\right) \\ & +\ (1-q)\ \ p\left(X\,|\,H_0\right)C_{00}\left(X,\hat{X}\right)\Big] \\ & + p\left(\eta_1\,|\,Y\right)\Big[q\,p\left(X\,|\,H_1\right)C_{11}\left(X,\hat{X}\right) \\ & +\ (1-q)\ \ p\left(X\,|\,H_0\right)C_{01}\left(X,\hat{X}\right)\Big]\Big\}\,dX\,dY\,. \end{aligned} \tag{6.10}$$

Let

$$r_{ij}\left(Y\right) = \int_{\Omega_x} C_{ij}\left(X,\hat{X}\right)p\left(X\,|\,H_i\right)p\left(Y\,|\,X\right)dX \tag{6.11}$$

denote a risk associated with the pair $\{H_i,\eta_j\}$ and the observation Y. Then, the combined Bayes risk follows

$$\begin{aligned} R = \int_{\Omega_y} & p\left(\eta_0\,|\,Y\right)\left[\,q\,r_{10}\left(Y\right)+(1-q)\,r_{00}\left(Y\right)\right] \\ & + p\left(\eta_1\,|\,Y\right)\left[\,q\,r_{11}\left(Y\right)+(1-q)\,r_{01}\left(Y\right)\right]dY\,. \end{aligned} \tag{6.12}$$

Since the detector's decision under a given observation is binary, *i.e.*, $p\left(\eta_j\,|\,Y\right)\in\{0,1\}$, for minimizing the combined risk we first evaluate the optimal estimator under each of the decisions, then the optimal decision rule is derived based on the optimal estimators $\hat{X}_0$, $\hat{X}_1$ to further minimize the combined risk. The two-stage minimization guaranties minimum combined risk [116]. The optimal *nonrandom* decision rule which minimizes the combined risk (6.12) is given by:

Decide η_1 (*i.e.*, $p\left(\eta_1\,|\,Y\right)=1$) if

$$q\left[r_{10}\left(Y\right)-r_{11}\left(Y\right)\right] \geq (1-q)\left[r_{01}\left(Y\right)-r_{00}\left(Y\right)\right]\ , \tag{6.13}$$

otherwise, decide η_0.

[2]Note that $X=0$ implies that H_0 is true and $X\neq 0$ implies H_1 so the sub-index i may seem to be redundant. However, this notation simplifies the subsequent formulations.

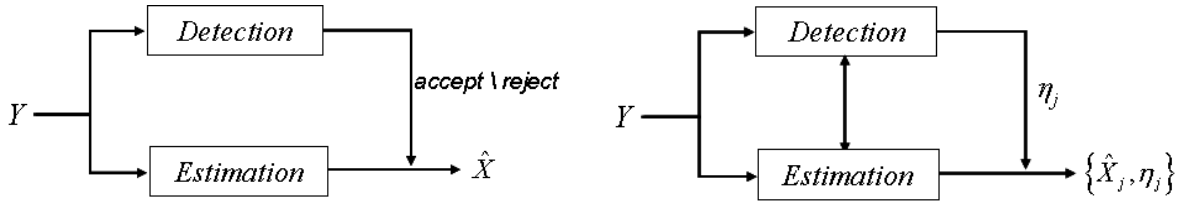


Figure 6.1: (a) Independent detection and estimation system; (b) strongly coupled detection and estimation system.

The optimal estimator under a decision η_j is obtained from (6.12) by

$$\arg\min_{\hat{X}_j} \{\, q\, r_{1j}\,(Y) + (1-q)\, r_{0j}\,(Y)\} \ . \tag{6.14}$$

Note that $r_{ij}\,(Y)$ depends on the estimate $\hat{X}_j$ through the cost function. Figure 6.1 shows a block diagram of the simultaneous detection and estimation scheme compared with an independent detection and estimation system. The standard, non-coupled detection and estimation system (a) consists of an estimator and a detector which independently chooses to accept or reject the estimator output. In the simultaneous detection and estimation scheme, the estimator is obtained by (6.14) and the interrelated decision rule (6.13) chooses the appropriate estimator, $\hat{X}_0$ or $\hat{X}_1$, for minimizing the combined Bayes risk. Since the risk $r_{ij}\,(Y)$ is a function of the signal estimate $\hat{X}_j$, the decision rule (6.13) requires knowledge of the estimator under any of its own decisions. Therefore, the arrow between the estimation and the detection blocks is unidirectional. It is important to note that the optimal estimator (6.14) minimizes the Bayes risk under any given decision rule, even if the detector is not optimal and/or is unknown to the estimator.

The cost function associated with the pair $\{H_i,\, \eta_j\}$ is generally defined by

$$C_{ij}\left(X, \hat{X}\right) = b_{ij}\, d_{ij}\left(X, \hat{X}\right) \ , \tag{6.15}$$

where $d_{ij}\left(X, \hat{X}\right)$ is an appropriate distortion measure and the cost parameters b_{ij} control the trade-off between the costs associated with the pairs $\{H_i, \eta_j\}$. That is, a high valued b_{01} raises the cost of a false alarm, (*i.e.,* decision of speech presence when speech is actually absent) which may result in residual musical noise. Similarly, b_{10} is associated with the cost of missed detection of a signal component, which may cause perceptual

signal distortion. Under a correct classification, normalized cost parameters are generally used, $b_{00} = b_{11} = 1$. However, $d_{ii}(\cdot,\cdot)$ is not necessarily zero since estimation errors are still possible even when there is no detection error.

Contrary to the approach in [115, 116, 139], we do not reject the signal estimator when a decision η_0 is made. Instead, we allow the estimator $\hat{X}_0 \neq 0$ to compensate for any detection errors and to reduce potential musical noise and audible distortions. Furthermore, when speech is indeed absent the distortion function is defined to allow some natural background noise level such that under H_0 the attenuation factor will be lower bounded by a constant gain floor $G_f << 1$ as proposed in [24, 29, 38, 70].

6.4 Quadratic spectral amplitude cost function

In this section, we derive a speech simultaneous detection and estimation scheme for a QSA cost function.

The distortion measure of the QSA cost function is defined by

$$d_{ij}\left(X,\hat{X}\right) = \begin{cases} \left(|X| - \left|\hat{X}_j\right|\right)^2 & i = 1\,, \\ \left(G_f\,|Y| - \left|\hat{X}_j\right|\right)^2 & i = 0\,, \end{cases} \tag{6.16}$$

and is related to the STSA suppression rule of Ephraim and Malah [33]. We assume that both X and D are statistically independent, zero-mean, complex-valued Gaussian random variables with variances λ_x and λ_d, respectively. Let $\xi \triangleq \lambda_x/\lambda_d$ denote the *a priori* SNR under hypothesis H_1, let $\gamma \triangleq |Y|^2/\lambda_d$ denote the *a posteriori* SNR and let $\upsilon \triangleq \gamma\,\xi/(1+\xi)$. For evaluating the optimal detector and estimator under the QSA cost function we denote by $X \triangleq a\,e^{j\alpha}$ and $Y \triangleq R\,e^{j\theta}$ the clean and noisy spectral coefficients, respectively, where $a = |X|$ and $R = |Y|$. Accordingly, the pdf of the speech expansion coefficient under H_1 satisfies

$$p\left(a,\alpha \mid H_1\right) = \frac{a}{\pi\lambda_x}\exp\left(-\frac{a^2}{\lambda_x}\right)\,. \tag{6.17}$$

The combined risk under the QSA cost function is independent of the signal phase nor the estimation phase. Therefore, we define $\hat{a}_j = \left|\hat{X}_j\right|$ as the estimated amplitude under

η_j. Substituting the QSA cost function into (6.14) we have

$$\hat{a}_j = \arg\min_{\hat{a}} \left\{ q\, b_{1j} \int_0^{\infty} \int_0^{2\pi} (a-\hat{a})^2\, p\,(a,\alpha \mid H_1)\, p\,(Y \mid a,\alpha)\, d\alpha\, da \right. \\ \left. + (1-q)\, b_{0j}\, (G_f\, R - \hat{a})^2\, p\,(Y \mid H_0) \right\}\,, \tag{6.18}$$

and by constraining the derivative according to $\hat{a}$ to equal zero, we obtain

$$\hat{a}_j\, [b_{1j}\, \Lambda\,(Y) + b_{0j}] = b_{1j}\, \Lambda\,(Y) \int_0^{\infty} \int_0^{2\pi} a\, p\,(a,\alpha \mid H_1)\, p\,(Y \mid a,\alpha)\, d\alpha\, da / p\,(Y \mid H_1) + b_{0j}\, G_f\, R \tag{6.19}$$

where $\Lambda\,(Y)$ is the generalized likelihood ratio defined by [33]

$$\begin{aligned} \Lambda\,(Y) &\triangleq \frac{q}{(1-q)} \frac{p\,(Y \mid H_1)}{p\,(Y \mid H_0)} \\ &= \frac{q}{(1-q)} \frac{e^{\upsilon}}{1+\xi}\,. \end{aligned} \tag{6.20}$$

Note that given the *a priori* speech presence probability, the generalized likelihood ratio is a function of the *a priori* and *a posteriori* SNRs, $\Lambda(\xi,\gamma)$. Using [33] we observe that

$$\begin{aligned} &\int_0^{\infty} \int_0^{2\pi} a\, p\,(a,\alpha \mid H_1)\, p\,(Y \mid a,\alpha)\, d\alpha\, da / p\,(Y \mid H_1) \\ &\quad = \frac{\sqrt{\pi\,\upsilon}}{2\gamma} \exp\left(-\frac{\upsilon}{2}\right) \left[(1+\upsilon)\, I_0\left(\frac{\upsilon}{2}\right) + \upsilon\, I_1\left(\frac{\upsilon}{2}\right)\right] R \\ &\quad \triangleq G_{STSA}\,(\xi,\gamma)\, R\,, \end{aligned} \tag{6.21}$$

where $I_\nu\,(\cdot)$ denotes the modified Bessel function of order ν.

Let

$$\phi_j\,(\xi,\gamma) \triangleq b_{1j}\, \Lambda\,(\xi,\gamma) + b_{0j}\,. \tag{6.22}$$

Then, by using the phase of the noisy signal [33] we obtain from (6.19) and (6.21) the optimal estimation under the decision η_j, $j \in \{0,1\}$:

$$\begin{aligned} \hat{X}_j &= [b_{1j}\, \Lambda\,(\xi,\gamma)\, G_{STSA}\,(\xi,\gamma) + b_{0j}\, G_f]\, \phi_j\,(\xi,\gamma)^{-1}\, Y \\ &\triangleq G_j\,(\xi,\gamma)\, Y\,. \end{aligned} \tag{6.23}$$

For evaluating the optimal decision rule we need to compute the risk $r_{ij}(Y)$. Under H_1

we obtain

$$\begin{aligned} r_{1j}(Y) &= \frac{b_{1j}}{\pi}\frac{\exp\left(-\frac{\gamma}{1+\xi}\right)}{1+\xi}\left\{G_j^2\,\gamma+\frac{\xi}{1+\xi}(1+\upsilon)-G_j\sqrt{\pi\,\upsilon}\exp\left(-\frac{\upsilon}{2}\right)\right. \\ &\qquad \left.\times\left[(1+\upsilon)\,I_0\left(\frac{\upsilon}{2}\right)+\upsilon\,I_1\left(\frac{\upsilon}{2}\right)\right]\right\} \\ &= \frac{b_{1j}}{\pi}\frac{\exp\left(-\frac{\gamma}{1+\xi}\right)}{1+\xi}\left\{G_j^2\,\gamma+\frac{\xi}{1+\xi}(1+\upsilon)-2\gamma\,G_j\,G_{STSA}\right\}, \end{aligned} \tag{6.24}$$

(see proof in the Appendix) where G_j holds for $G_j(\xi,\gamma)$, the gain function under the QSA cost function and the decision η_j which is defined in (6.23), and G_{STSA} holds for $G_{STSA}(\xi,\gamma)$ which is defined in (6.21).

For deriving the risk under H_0, $r_{0j}(Y)$, we observe $p(X_R,X_I\,|\,H_0)=\delta(X_R,X_I)$. Consequently,

$$\begin{aligned} r_{0j}(Y) &= b_{0j}\int\int_{-\infty}^{\infty}\left\{\left[G_j(\xi,\gamma)-G_f\right]^2|Y|^2\right\}p(X_R,X_I\,|\,H_0)\,p(Y\,|\,X_R,X_I)\,dX_R\,dX_I \\ &= \frac{b_{0j}}{\pi}\left[G_j(\xi,\gamma)-G_f\right]^2\gamma\,e^{-\gamma}. \end{aligned} \tag{6.25}$$

Substituting (6.24) and (6.25) into (6.13), we obtain the optimal decision rule under the QSA cost function:

$$\begin{aligned} &\Lambda(\xi,\gamma)\left\{b_{10}\,G_0^2-G_1^2+\frac{\xi}{(1+\xi)\,\gamma}(1+\upsilon)(b_{10}-1)+2(G_1-b_{10}\,G_0)\,G_{STSA}\right\} \\ &\qquad \underset{\eta_0}{\overset{\eta_1}{\gtrless}} b_{01}(G_1-G_f)^2-(G_0-G_f)^2. \end{aligned} \tag{6.26}$$

To conclude the above results, simultaneous detection and estimation from noisy observations requires (i) calculating the gain factor under any of the decisions using (6.23), and (ii) finding the optimal decision η_j using (6.26). The corresponding signal estimate is obtained by applying the gain G_j to the noisy observation.

Figure 6.2 demonstrates attenuation curves under QSA cost function as a function of the *instantaneous* SNR defined by $\gamma-1$, for several *a priori* SNRs, using the parameters $q=0.8$, (as proposed in [33]) $G_f=-25$ dB and cost parameters $b_{01}=5$ and $b_{10}=1.1$. The gains G_1 (dashed line), G_0 (dotted line) and the total detection and estimation system gain (solid line) are compared to the STSA gain under signal presence uncertainty of Ephraim and Malah [33] (dashed-dotted line). The *a priori* SNRs range from -15 dB to 15 dB. Not only that the cost parameters shape the STSA gain curve, when combined

with the detector the proposed method provides a significant non-continuous modification of the standard STSA estimator. For example, for *a priori* SNRs of $\xi = -5$ and $\xi = 15$ dB, as shown in Figure 6.2(b) and (d) respectively, as long as the instantaneous SNR is higher than about -2 dB (for $\xi = -5$ dB) or -5 dB (for $\xi = 15$ dB), the detector decision is η_1, while for lower instantaneous SNRs, the detector decision is η_0. Note that if an ideal detector for the speech coefficients would be available, a more significantly non-continuous gain would be desired to block the noise-only coefficients. However, in the proposed simultaneous detection and estimation approach the detector is not ideal but optimized to minimize the combined risk and the non-continuity of the system gain depends on the chosen cost parameters as well as on the gain floor. As shown in our experimental results, this non-continues gain function may yield greater noise reduction with slightly higher level of musicality, while not degrading speech quality.

It is of interest to examine the asymptotic behavior of the estimator (6.23) under each of the decisions. When the cost parameter associated with false alarm is much smaller than the generalized likelihood ratio, *i.e.*, $b_{01} << \Lambda(\xi,\gamma)$, the spectral gain of the estimator under the decision η_1 is $G_1(\xi,\gamma) \cong G_{STSA}(\xi,\gamma)$, which is optimal when the signal is surely present. However, if $b_{01} >> \Lambda(\xi,\gamma)$, the spectral gain under η_1 needs to compensate the possibility of a high-cost false-decision made by the detector and thus $G_1(\xi,\gamma) \cong G_f$. On the other hand, if the cost parameter associated with missed detection is small and we have $b_{10} << \Lambda(\xi,\gamma)^{-1}$, then $G_0(\xi,\gamma) \cong G_f$ (*i.e.*, estimation where speech is surely absent) but under $b_{10} >> \Lambda(\xi,\gamma)^{-1}$, in order to overcome the high cost related to missed detection, we have $G_0(\xi,\gamma) \cong G_{STSA}(\xi)$.

Recall that

$$\frac{\Lambda(\xi,\gamma)}{1+\Lambda(\xi,\gamma)} = p(H_1 \mid Y) \tag{6.27}$$

is the *a posteriori* probability for speech presence [33], it can be shown that the proposed estimator (6.23) generalizes the well-known STSA estimator. For the case of $b_{ij} = 1\ \forall i,j$ we have

$$\begin{aligned} \hat{X}_0 &= [\,p(H_1 \mid Y)\, G_{STSA}(\xi,\gamma) + (1 - p(H_1 \mid Y))\, G_f]\, Y \\ &= \hat{X}_1\,. \end{aligned} \tag{6.28}$$

In that case the detection operation is not required since the estimation is independent of

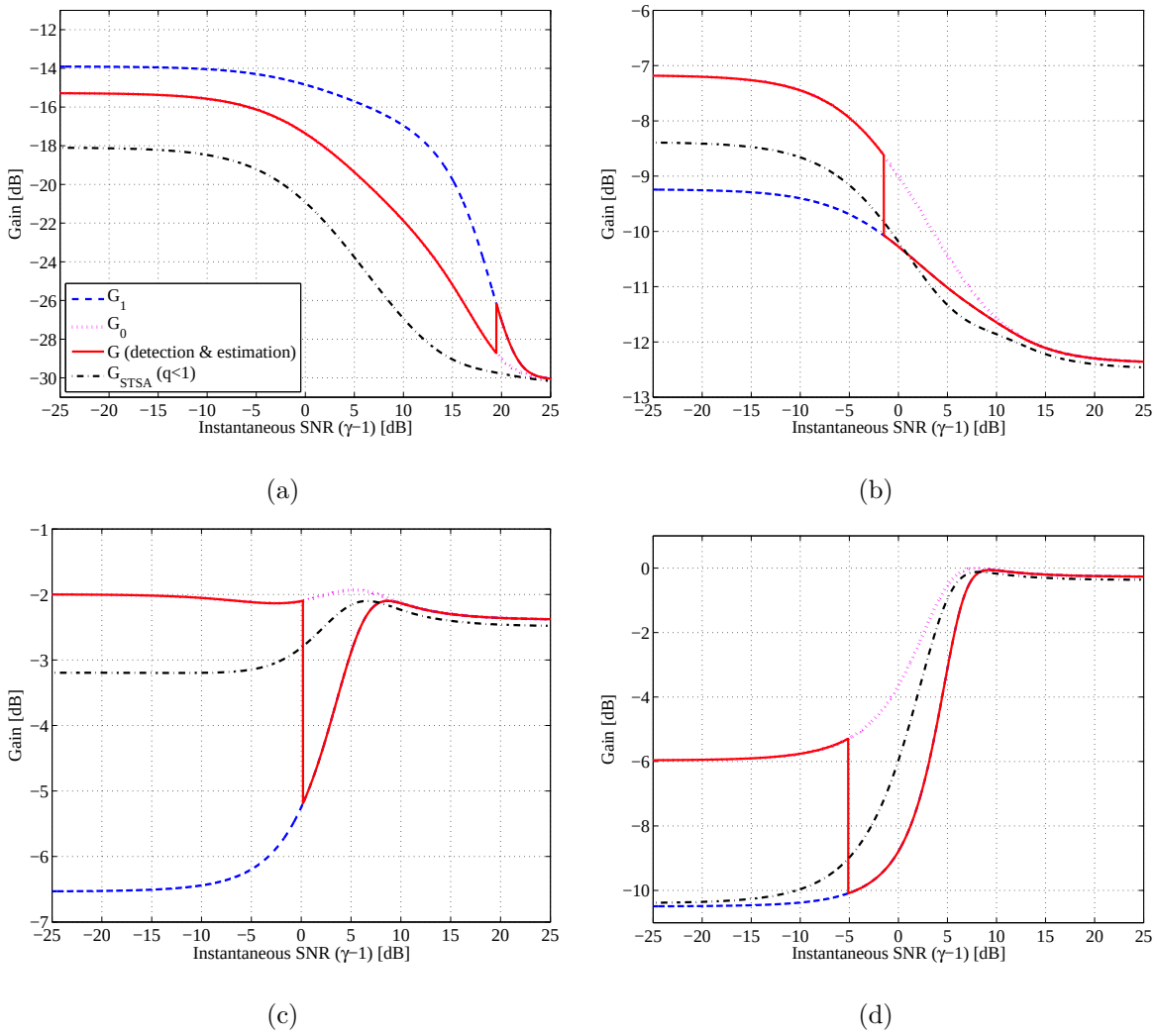


Figure 6.2: Gain curves of G_1 (dashed line), G_0 (dotted line) and the total detection and estimation system gain curve (solid line), compared with the STSA gain under signal presence uncertainty (dashed-dotted line). The *a priori* SNRs are (a) $\xi = -15$ dB, (b) $\xi = -5$ dB, (c) $\xi = 5$ dB and (d) $\xi = 15$ dB.

the decision rule. If we also set G_f to zero, the estimation reduces to the STSA suppression rule under signal presence uncertainty [33].

The simultaneous detection and estimation approach requires the calculation of two gain functions, $G_0(\xi,\gamma)$ and $G_1(\xi,\gamma)$, and the decision rule. However, as can be seen from (6.23), both $G_0(\xi,\gamma)$ and $G_1(\xi,\gamma)$ are linear functions of $G_{STSA}(\xi,\gamma)$ and the generalized likelihood ratio $\Lambda(\xi,\gamma)$. In addition, the decision rule (6.26) requires the calculation of a second-order polynomial. Therefor, the additional complexity of the simultaneous detection and estimation approach is insignificant compared to the STSA estimator [33], which also requires the calculation of the gain function $G_{STSA}(\xi,\gamma)$ (6.21) and the generalized likelihood function (6.31).

6.5 Relation to spectral subtraction

The general formulation of the spectral subtraction approach assumes a spectral estimator which can be written as [29, 30]

$$\hat{X}_{\ell k} = \max\left\{\left(|Y_{\ell k}|^{\tau} - \mu E\left[|D_{\ell k}|^{\tau}\right]\right)^{\frac{1}{\tau}}, \beta E\left[|D_{\ell k}|^{\tau}\right]^{\frac{1}{\tau}}\right\}\frac{Y_{\ell k}}{|Y_{\ell k}|} \tag{6.29}$$

where $E\left[|D_{\ell k}|^{\tau}\right]$ is the τ-order moment of the noise spectral coefficient, $\mu \geq 1$ represents an over-subtraction factor, and $0 < \beta << 1$ represents spectral floor factor. Boll [30] considered $\tau = 1$ while Berouti *et al.* [29] used $\tau = 2$. McAulay and Malpass [32] showed that under a Gaussian statistical model, spectral subtraction with $\tau = 2$, $\mu = 1$ and $\beta = 0$ yields a maximum-likelihood estimator for the speech spectral variance.

The spectral subtraction scheme (6.29) classifies high-energy time-frequency bins as active speech bins, and only in these bins the signal is estimated. Low-energy bins below a given threshold are classified as noise-only bins, and set to some background noise level for reducing the residual musical noise. Consequently, low-energy bins that contain the speech signal are not detected, while noise peaks are detected as speech bins. When the over-subtraction factor μ is increased, fewer noise peaks are detected as speech and therefore the residual musical noise is reduced at the expanse of deterioration of speech quality. The spectral floor $\beta E\left[|D_{\ell k}|^{\tau}\right]^{1/\tau}$ "fills-in" the valleys of the residual noise, which yields a more natural noise with less annoying musicality [29]. However, a large β reduces

the background noise suppression. Further reduction of the musical noise may be achieved by local smoothing of the noisy spectral values prior to noise subtraction. As a result, noise peaks are attenuated and the spectral estimation error can be reduced [30]. However, as the speech signal is highly nonstationary, its intelligibility may be dramatically decreased when the smoothing parameter increases.

The classical spectral subtraction approach heuristically combines a detector and an estimator for the speech spectral coefficients while the parameters μ, β and the smoothing length control the trade-off between the residual musical noise and the speech quality. In the proposed simultaneous detection and estimation approach, the detector is optimally designed jointly with the estimator. The residual noise musicality is controlled by both the spectral gain floor G_f which bounds the attenuation and the false-alarm cost parameter b_{01}. A high-valued false-alarm cost parameter (with relation to the generalized likelihood ratio) reduces the estimation gain under η_1, which compensates for a false-detection. The amount of speech distortion is affected by the missed detection parameter b_{10}, which increases the estimation gain under η_0. Since the decision rule depends on both parameters as well as on the gain floor, it is the combination of the three parameters that control the trade-offs between noise reduction and speech distortion.

The different behaviors of the spectral subtraction and the simultaneous detection and estimation approach are illustrated in Figures 6.3 and 6.4. The signals in the time domain are shown in Figure 6.3. The clean signal is a sinusoidal wave which is active only in a specific time interval and the noisy signal contains white Gaussian noise with SNR of 5 dB. The noisy signal is transformed into the STFT domain using half-overlapping Hamming windows of 256 taps. The signal enhanced by spectral subtraction with $\tau = 2$, $\mu = 1$ and $\beta = 0.2$ is shown in Figure 6.3(c) and the signal enhanced by using the proposed algorithm is shown in Figure 6.3(d) with $b_{01} = 3$, $b_{10} = 5$, $G_f = -20$ dB and $q = 0.8$. The *a priori* SNR needed for the simultaneous detection and estimation approach is estimated using the decision-directed approach as will be defined in (6.30), with a weighting factor $\alpha = 0.92$ and $\xi_{min} = -20$ dB as the lower bound for the *a priori* SNR, while the variance of the background noise coefficients is evaluated from the noise signal (for both algorithms). The amplitudes of the signals in the STFT domain (at the specific frequency band of the desired signal's frequency) are shown in Figure

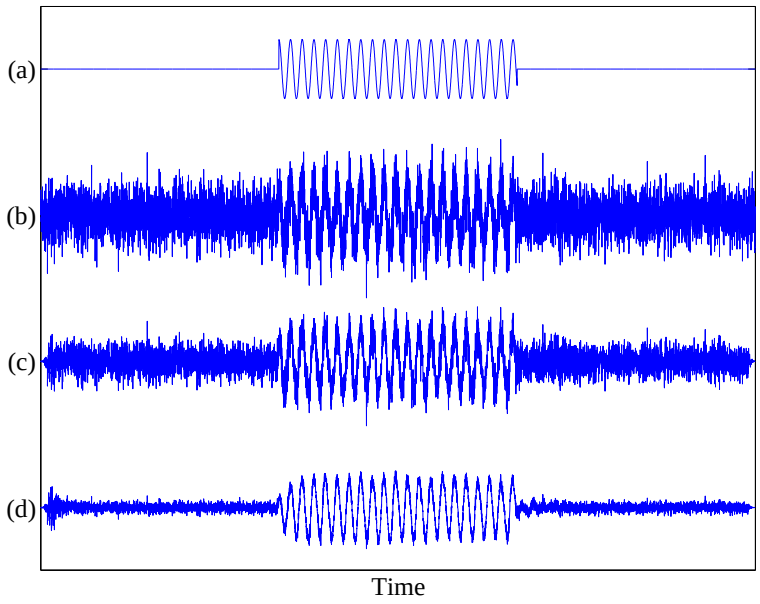


Figure 6.3: Signals in the time domain. (a) Clean sinusoidal signal; (b) noisy signal; (c) enhanced signal obtained by using the spectral-subtraction estimator; (d) enhanced signal obtained by using the detection and estimation approach.

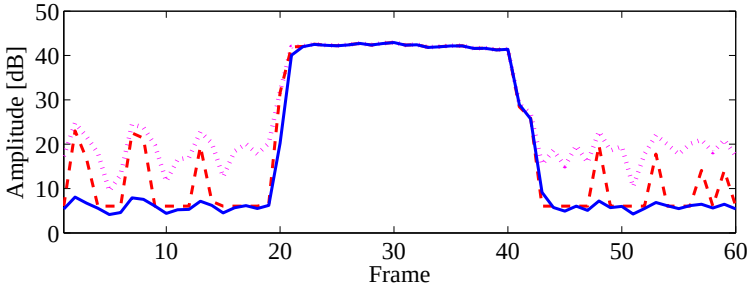


Figure 6.4: Amplitudes of the STFT coefficients along the time-trajectory corresponding to the frequency of the sinusoidal signal: noisy signal (dotted line), spectral subtraction (dashed line), and simultaneous detection and estimation (solid line).

6.4. It can be seen that when the desired signal is absent, high-energy noise components are falsely detected by the spectral subtraction algorithm which potentially results in an annoying musical noise. The detection and estimation algorithm results in a higher attenuation of the noise peaks and smoother and more natural background noise while not increasing the audible distortion in the enhanced signal. Furthermore, it may seem from Figure 6.4 that when the desired signal is active and the instantaneous SNR is high, both algorithms imply similar results. However, in time frames where the desired signal is present, the spectral subtraction approach results in higher residual noise in frequencies where the signal is absent or of low SNR. Therefore, the enhanced signal using the spectral subtraction approach is inferior to the enhanced signal using the detection and estimation approach even in time intervals where the signal is present, as can be seen from Figures 6.3(c) and (d).

6.6 A priori SNR estimation

Speech enhancement in the STFT domain generally relies on an estimation-only approach under signal presence uncertainty *e.g.*, [32, 33, 38]. The *a priori* SNR is often estimated by using the decision-directed approach [33]. Accordingly, in each time-frequency bin we compute

$$\hat{\xi}_{\ell k} = \max\left\{\alpha\, G^2\left(\hat{\xi}_{\ell-1,k}, \gamma_{\ell-1,k}\right)\gamma_{\ell-1,k}\,(1-\alpha)\,(\gamma_{\ell k}-1)\,,\,\xi_{\min}\right\} \tag{6.30}$$

where α $(0 \leq \alpha \leq 1)$ is a weighting factor that controls the trade-off between noise reduction and transient distortion introduced into the signal, and $\xi_{\min}$ is a lower bound for the *a priori* SNR which is necessary for reducing the residual musical noise in the enhanced signal [33, 70]. Since the *a priori* SNR is defined under the assumption that $H_1^{\ell k}$ is true, it is proposed in [38] to replace the gain G in (6.30) by G_{H_1} which represents the spectral gain when the signal is surely present (*i.e.*, $q = 1$). Increasing the value of α results in a greater reduction of the musical noise phenomena, at the expense of further attenuation of transient speech components (*e.g.*, speech onsets) [70]. By using the proposed approach with high cost for false speech detection, the musical noise can be reduced without increasing the value of α, which enables rapid changes in the *a priori* SNR estimate. The lower bound for the *a priori* SNR is related to the spectral gain floor

G_f since both imply a lower bound on the spectral gain. The latter parameter is used to evaluate both the optimal detector and estimator while taking into account the desired residual noise level.

The decision-directed estimator is widely used, but is not suitable for transient noise environments, since a high-energy noise burst may yield an instantaneous increase in the *a posteriori* SNR and a corresponding increase in $\hat{\xi}_{\ell k}$ as can be seen from (6.30). The spectral gain would then be higher than the desired value, and the transient noise component would not be sufficiently attenuated. Let $\hat{\lambda}^s_{d_{\ell k}}$ denote the estimated spectral variance of the stationary noise component and let $\hat{\lambda}^t_{d_{\ell k}}$ denote the estimated spectral variance of the transient component. The former may be practically estimated by using the improved minima-controlled recursive averaging (IMCRA) algorithm [38, 71] or by using the minimum-statistics approach [72], while $\lambda^t_{d_{\ell k}}$ may be evaluated based on a training phase as assumed in [140]. The total variance of the noise component is $\hat{\lambda}_{d_{\ell k}} = \hat{\lambda}^s_{d_{\ell k}} + \hat{\lambda}^t_{d_{\ell k}}$. Note that $\lambda^t_{d_{\ell k}} = 0$ in time-frequency bins where the transient noise is inactive. Since the *a priori* SNR is highly dependent on the noise variance, we first estimate the speech spectral variance by

$$\hat{\lambda}_{x_{\ell k}} = \max\left\{\alpha\, G^2_{H_1}\left(\hat{\xi}_{\ell-1,k}, \gamma_{\ell-1,k}\right)|Y_{\ell-1,k}|^2 + (1-\alpha)\left(|Y_{\ell k}|^2 - \hat{\lambda}_{d_{\ell k}}\right),\ \lambda_{\min}\right\} \tag{6.31}$$

where $\lambda_{\min} = \xi_{\min}\,\hat{\lambda}^s_{d_{\ell k}}$. Then, the *a priori* SNR is evaluated by $\hat{\xi}_{\ell k} = \hat{\lambda}_{x_{\ell k}}/\hat{\lambda}_{d_{\ell k}}$. It is straightforward to show that in a stationary noise environment the proposed *a priori* SNR estimator reduces to the decision-directed estimator (6.30), with G_{H_1} substituting G. However, under the presence of a transient noise component, the proposed method yields a lower *a priori* SNR estimate, which enables higher attenuation of the high-energy transient noisy component. Furthermore, to allow further reduction of the transient noise component to the level of the residual stationary noise, we modify the gain floor by $\tilde{G}_f = G_f\,\hat{\lambda}^s_{d_{\ell k}}/\hat{\lambda}_{d_{\ell k}}$ as proposed in [141].

The different behaviors under transient noise conditions of the proposed modified decision-directed *a priori* SNR estimator and the decision-directed estimator as proposed in [38] are illustrated in Figures 6.5 and 6.6. Figure 6.5 shows the signals in the time domain: the analyzed signal contains a sinusoidal wave which is active in only two specific segments. The noisy signal contains both additive white Gaussian noise with 5 dB SNR

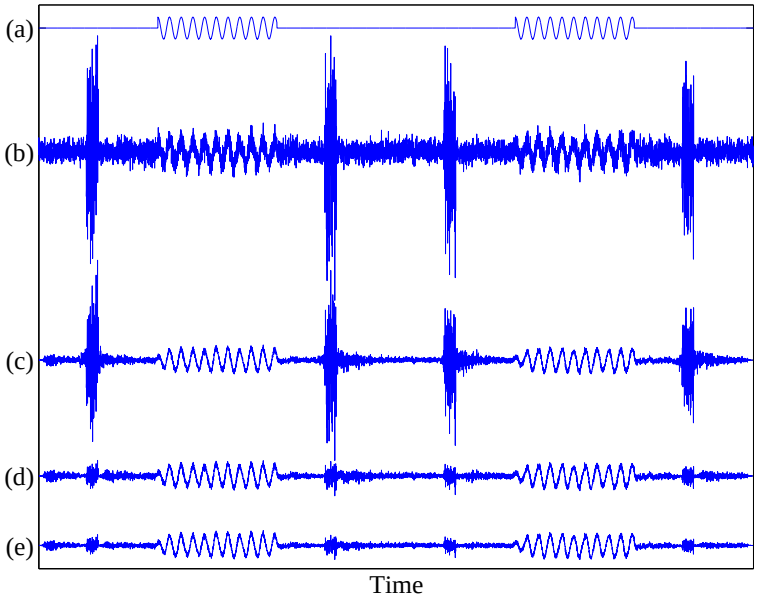


Figure 6.5: Signals in the time domain. (a) Clean sinusoidal signal; (b) noisy signal with both stationary and transient components; (c) enhanced signal obtained by using the STSA and the decision-directed estimators; (d) enhanced signal obtained by using the STSA and the modified *a priori* SNR estimators ;(e) enhanced signal obtained by using the detection and estimation approach and the modified *a priori* SNR estimator.

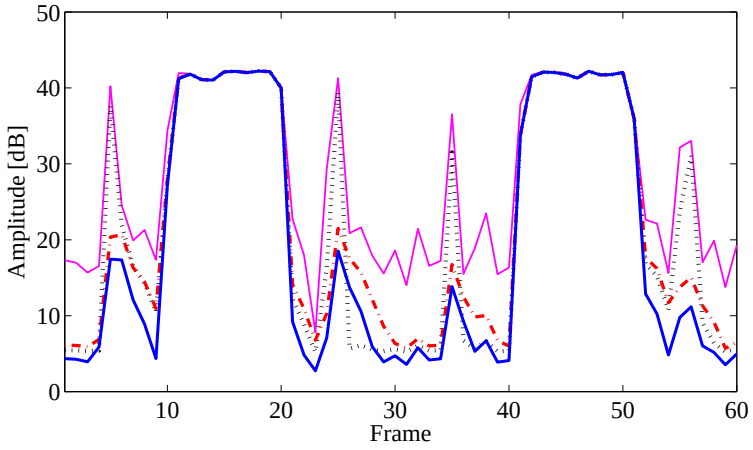


Figure 6.6: Amplitudes of the STFT coefficients along time-trajectory corresponding to the frequency of the sinusoidal signal: noisy signal (light solid line), STSA with decision-directed estimation (dotted line), STSA with the modified *a priori* SNR estimator (dashed-dotted line) and simultaneous detection and estimation with the modified *a priori* SNR estimator (dark solid line).

and high-energy transient noise components. The signal enhanced by using the decision-directed estimator and the STSA suppression rule is shown in Figure 6.5(c). The signal enhanced by using the modified *a priori* SNR estimator and the STSA suppression rule is shown in Figure 6.5(d), and the result obtained by using the proposed modified *a priori* SNR estimation with the detection and estimation approach is shown in Figure 6.5(d) (using the same parameters as in the previous section). Both the decision-directed estimator and the modified *a priori* SNR estimator are applied with $\alpha = 0.98$ and $\xi_{\min} = -20$ dB. Clearly, in stationary noise intervals, and where the SNR is high, similar results are obtained by both *a priori* SNR estimators. However, the proposed modified *a priori* SNR estimator obtain higher attenuation of the transient noise, whether it is incorporated with the STSA or the simultaneous detection and estimation approach. Figure 6.6 shows the amplitudes of the STFT coefficients of the noisy and enhanced signals at the frequency band which contains the desired sinusoidal component. Accordingly, the modified *a priori* SNR estimator enables a greater reduction of the background noise, particularly transient noise components. Moreover, it can be seen that using the simultaneous detection and estimation yields better attenuation of both the stationary and background noise compared to the STSA estimator, even while using the same *a priori* SNR estimator.

6.7 Experimental results

In our experimental study, we first evaluate the detection and estimation approach compared with the STSA suppression rule under a stationary noise environment. Then, we consider the problem of hands-free communication in an emergency car, and demonstrate the advantage of the modified *a priori* SNR estimator together with the simultaneous detection and estimation approach under transient noise environment. Speech signals are taken from the TIMIT database [142], sampled at 16 kHz and degraded by additive noise. The test signals include 16 speech utterances from 16 different speakers, half male half female. The noisy signals are transformed into the STFT domain using half-overlapping Hamming windows of 32 msec length, and the background-noise spectrum is estimated by using the IMCRA algorithm (for all the considered enhancement algorithms) [38, 71]. The performance evaluation in our study includes objective quality measures, a subjec-

Table 6.1: Segmental SNR and Log Spectral Distortion Obtained by Using Either the Simultaneous Detection and Estimation Approach or the STSA Estimator in Stationary Noise Environment.

Input SNR	Input Signal		Detection & Estimation		STSA ($\alpha = 0.98$)		STSA ($\alpha = 0.92$)	
dB	SegSNR	LSD	SegSNR	LSD	SegSNR	LSD	SegSNR	LSD
-5	-6.801	20.897	1.255	7.462	0.085	9.556	-0.684	10.875
0	-3.797	16.405	4.136	5.242	3.169	6.386	2.692	7.391
5	0.013	12.130	5.98	3.887	5.266	4.238	5.110	4.747
10	4.380	8.194	6.27	3.143	5.93	3.167	6.014	3.157

tive study of spectrograms and informal listening tests. The first quality measure is the segmental SNR defined by [143]

$$\text{SegSNR} = \frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}}\mathcal{T}\left\{10\log_{10}\frac{\sum_{n=0}^{K-1}x^2\left(n+\ell\,K/2\right)}{\sum_{n=0}^{K-1}\left[x\left(n+\ell\,K/2\right)-\hat{x}\left(n+\ell\,K/2\right)\right]^2}\right\}, \qquad (6.32)$$

where $\mathcal{L}$ represents the set of frames which contain speech, $|\mathcal{L}|$ denotes the number of elements in $\mathcal{L}$, $K = 512$ is the number of samples per frame and the operator $\mathcal{T}$ confines the SNR at each frame to a perceptually meaningful range between -10 dB and 35 dB. The second quality measure is log-spectral distortion (LSD) which is defined by

$$\text{LSD} = \frac{1}{L}\sum_{\ell=0}^{L-1}\left\{\frac{1}{K/2+1}\sum_{k=0}^{K/2}\left[10\log_{10}\mathcal{C}X_{\ell k}-10\log_{10}\mathcal{C}\hat{X}_{\ell k}\right]^2\right\}^{\frac{1}{2}}, \qquad (6.33)$$

where $\mathcal{C}X \triangleq \max\left\{|X|^2,\epsilon\right\}$ is a spectral power clipped such that the log-spectrum dynamic range is confined to about 50 dB, that is, $\epsilon = 10^{-50/10}\cdot\max_{\ell,k}\left\{|X_{\ell k}|^2\right\}$. The third quality measure (used in Section 6.7-B) is the perceptual evaluation of speech quality (PESQ) score [144].

6.7.1 Comparison with the STSA estimator

In this section, the suppression rule results from the proposed simultaneous detection and estimation approach is compared to the STSA estimation [33] for stationary white Gaussian noise with SNRs in the range $[-5, 10]$ dB. For both algorithms the *a priori* SNR is estimated by the decision-directed approach (6.30) with $\xi_{min} = -15$ dB, and the *a priori*

speech presence probability is $q_{\ell k} = 0.8$, as proposed in [33]. For the STSA estimator a decision-directed estimation [38] with $\alpha = 0.98$ reduces the residual musical noise but generally implies transient distortion of the speech signal [33, 70]. However, the inherent detector obtained by the simultaneous detection and estimation approach may improve the residual noise reduction and therefore a lower weighting factor α may be used to allow lower speech distortion. Indeed, we have found out that for the simultaneous detection and estimation approach $\alpha = 0.92$ implies better results, while for the STSA algorithm, better results are achieved with $\alpha = 0.98$. The cost parameters for the simultaneous detection and estimation should be chosen according to the system specification, *i.e.*, whether the quality of the speech signal or the amount of noise reduction is of higher importance. Table 6.1 summarizes the average segmental SNR and LSD for these two enhancement algorithms, with cost parameters $b_{01} = 10$ and $b_{10} = 2$, and $G_f = -15$ dB for the simultaneous detection and estimation algorithm. The results for the STSA algorithm are presented for $\alpha = 0.98$ as well as for $\alpha = 0.92$ (note that for the STSA estimator $G_f = 0$ is considered as originally proposed). It shows that the simultaneous detection and estimation yields improved segmental SNR and LSD, while a greater improvement is achieved for lower input SNR. Informal subjective listening tests and inspection of spectrograms demonstrate improved speech quality with higher attenuation of the background noise. However, since the weighting factor used for the *a priori* SNR estimate is lower, and the gain function is discontinuous, the residual noise resulting from the simultaneous detection and estimation algorithm is slightly more musical than that resulting from the STSA algorithm (examples are available online [145]).

6.7.2 Speech enhancement under nonstationary noise environment

In this section, we demonstrate the potential advantage of the simultaneous detection and estimation approach with the proposed *a priori* SNR estimator under transient noise. We consider a hands-free communication in an emergency car (police car, ambulance *etc.*) where the engine noise is assumed quasi-stationary. However, activating the emergency siren significantly degrades the perceptual quality and intelligibility of the speech signal,

Table 6.2: Segmental SNR, Log Spectral Distortion and PESQ Score Under Transient Noise.

	SegSNR	LSD	PESQ
Input Signal	-6.703	6.587	2.017
OM-LSA	-4.94	5.338	2.141
STSA	4.502	3.580	2.839
Detection and estimation			
$b_{01} = b_{10} = 1.5$	5.761	3.236	3.072
Detection and estimation			
$b_{01} = b_{10} = 5$	6.506	3.141	3.071

since its energy is much higher than that of the speech signal. The sound generation in a siren is nonlinear, which produces harmonics not present in the original signal (siren source signal), as can be seen in Figure 6.7(b). However, using the available siren source signal, a reliable indicator in the time-frequency domain for the presence of siren noise, and an estimate for the variance of the transient noise, $\lambda^t_{d\ell k}$, may be designed in a training phase. Note that standard echo-cancellation algorithms are not suitable for eliminating noise generated by nonlinear systems and nonlinear algorithms may be required (*e.g.*, [146,147]).

The proposed approach is compared with the STSA algorithm [33] and the OM-LSA algorithm [38]. The speech presence probability required for the OM-LSA estimator as well as for the simultaneous detection and estimation approach is estimated as proposed in [38], while for the STSA estimator $\hat{q}_{\ell k} = 0.8$ is used as originally proposed in [33]. However, since the *a priori* SNR estimate has a major importance under transient noise, the proposed modified decision-directed estimator is applied both for the simultaneous detection and estimation approach and for the STSA algorithm with $\xi_{min} = -20$ dB. For the simultaneous detection and estimation algorithm $\alpha = 0.92$ is used while for the STSA algorithm $\alpha = 0.98$ (as shown in Section 6.7.1 to be more appropriate for the STSA estimator). For the OM-LSA algorithm, the decision-directed estimator with $\alpha = 0.92$ is implemented as specified in [38] and the gain floor is $G_f = -20$ dB. Figure 6.7 shows waveforms and spectrograms of a clean signal, noisy signal and enhanced signals. The noisy signal contains engine car noise with 0 dB SNR and additional siren noise with -1 dB SNR, such that the total SNR is about -3 dB. The speech enhanced by using the OM-LSA

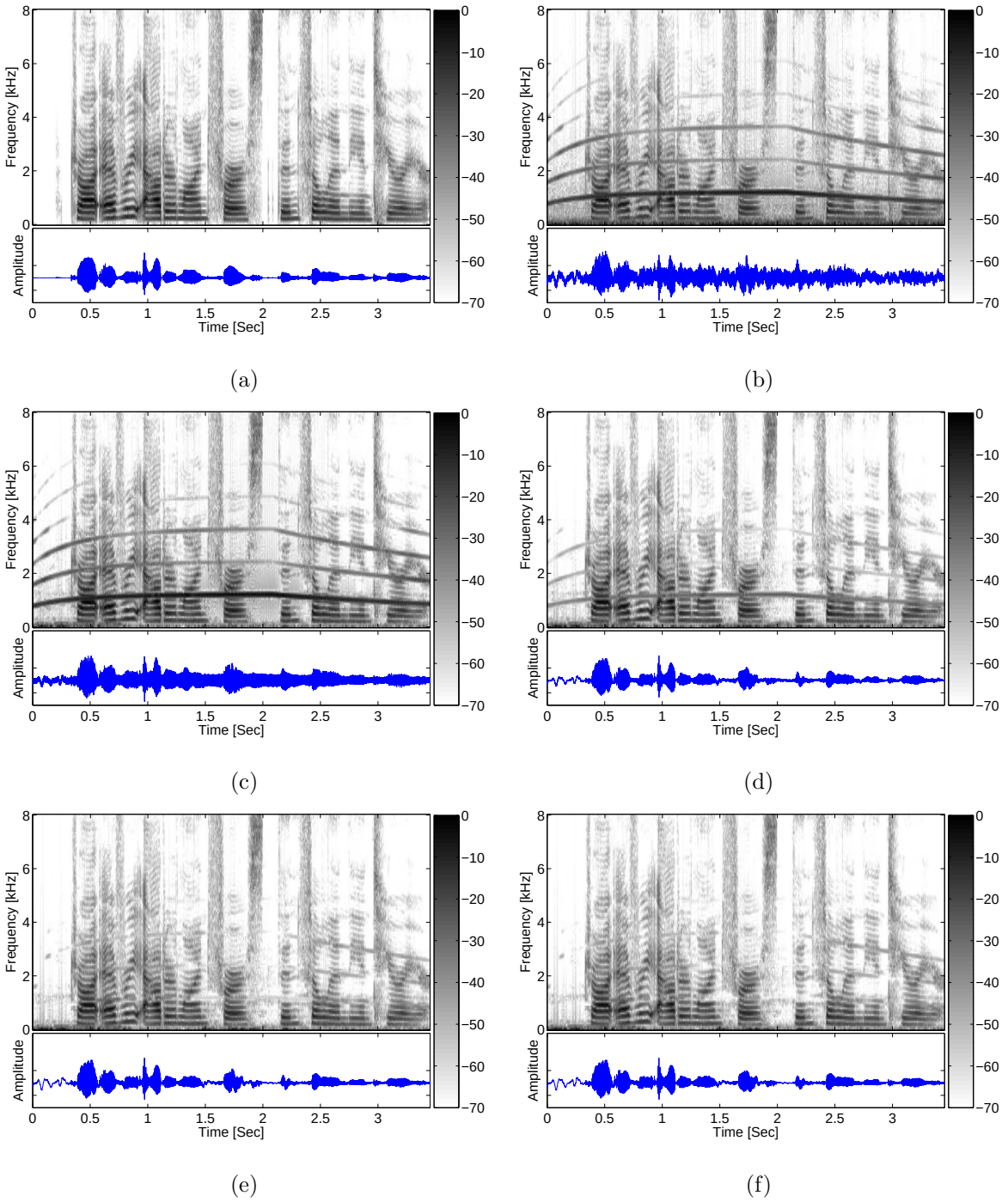


Figure 6.7: Speech spectrograms (in dB) and waveforms. (a) Clean speech signal: "Draw every outer line first, then fill in the interior"; (b) speech degraded by engine car noise and siren noise with SNR of -3 dB; (c) speech enhanced by using the OM-LSA estimator; (d) speech enhanced by using the STSA estimator (together with the modified *a priori* SNR estimator); (e) speech enhanced by using the simultaneous detection and estimation approach with $b_{01} = b_{10} = 1.5$; (f) speech enhanced by using the simultaneous detection and estimation approach with $b_{01} = b_{10} = 5$.

algorithm and the STSA algorithm are shown in Figures 6.7(c) and (d), respectively. The signal enhanced by using the simultaneous detection and estimation approach is shown in Figures 6.7(e) and (f) with $b_{01} = b_{10} = 1.5$ and $b_{01} = b_{10} = 5$, respectively, and a gain floor of $G_f = -20$ dB. It can be seen that compared with the decision-directed-based OM-LSA algorithm, the modified *a priori* SNR estimator substantially contributes to the transient noise reduction, whether it is integrated with the simultaneous detection and estimation approach or with the STSA algorithm. However, the simultaneous detection and estimation approach which is combined with adapted speech presence probability and gain floor yields greater reduction of transient noise without affecting the quality of the enhanced speech signal. Averaged quality measures for the whole set of tested utterances are summarized in Table 6.2, for the same noise conditions. The results demonstrate improved speech quality obtained by using the modified *a priori* SNR estimator either while combined with the STSA or the simultaneous detection and estimation approach, applying the detection and estimation approach introduced additional improvement to the enhanced signal. Subjective listening tests confirm that the speech quality improvement achieved by using the proposed method is perceptually substantial (audio files are available online [145]).

6.8 Conclusions

We have presented a novel formulation of the single-channel speech enhancement problem in the time-frequency domain. Our formulation relies on coupled operations of detection and estimation in the STFT domain, and a cost function that combines both the estimation and detection errors. A detector for the speech coefficients and a corresponding estimator for their values are jointly designed to minimize a combined Bayes risk. In addition, cost parameters enable to control the trade-off between speech quality, noise reduction and residual musical noise. The proposed method generalizes the traditional spectral enhancement approach which considers estimation-only under signal presence uncertainty. In addition we propose a modified decision-directed *a priori* SNR estimator which is adapted to transient noise environment. Experimental results show greater noise reduction with improved speech quality when compared with the STSA suppression rules

under stationary noise. Furthermore, it is demonstrated that under transient noise environment, greater reduction of transient noise components may be achieved by exploiting reliable information for the *a priori* SNR estimation with simultaneous detection and estimation approach.

6.A Risk derivation

In this appendix we derive the risk $r_{1j}(Y)$. Under $\{H_1, \eta_j\}$ we obtain

$$r_{1j}(Y) = b_{1j} \int_0^\infty \int_0^{2\pi} (a - G_j R)^2 \, p(a, \alpha \mid H_1) \, p(Y \mid a, \alpha) \, d\alpha \, da \,, \tag{6.34}$$

and the multiplication of the two pdf's implies

$$p(a, \alpha \mid H_1) \, p(Y \mid a, \alpha) = \frac{a}{\pi^2 \lambda_x \lambda_d} \exp\left\{ - \left(\gamma + \frac{a^2}{\lambda} - \frac{2R\, a \cos(\alpha - \theta)}{\lambda_d} \right) \right\} \,, \tag{6.35}$$

where $\lambda \triangleq (1/\lambda_x + 1/\lambda_d)^{-1}$. Integrating (6.34) with regard to the phase variable we obtain [148, eq. 3.339, 8.406.3]

$$\int_0^{2\pi} \exp\left\{ \frac{2R\, a \cos(\alpha - \theta)}{\lambda_d} \right\} d\alpha = 2\pi \, J_0\left(i \frac{2R}{\lambda_d} a \right) \,, \tag{6.36}$$

where $J_0(\cdot)$ denotes the Bessel function of order zero. Note that in this appendix $i \triangleq \sqrt{-1}$. Using [149, eq. 13.3.1, 2] we have

$$\int_0^\infty a \exp\left(-\frac{a^2}{\lambda} \right) J_0\left(i \frac{2R}{\lambda_d} a \right) da = \frac{\lambda}{2} e^{\upsilon} \,, \tag{6.37}$$

and

$$\int_0^\infty a^2 \exp\left(-\frac{a^2}{\lambda} \right) J_0\left(i \frac{2R}{\lambda_d} a \right) da = \frac{\lambda^{1.5} \Gamma(1.5)}{2\Gamma(1)} \, {}_1F_1(1.5; 1; \upsilon) \,, \tag{6.38}$$

where $\Gamma(\cdot)$ denotes the Gamma function with $\Gamma(1) = 1$ and $\Gamma(1.5) = \sqrt{\pi}/2$, and ${}_1F_1(a; b; x)$ is the confluent hypergeometric function [150, eq. A.1.31.c]

$${}_1F_1(1.5; 1; \upsilon) = e^{\frac{\upsilon}{2}} \left[(1 + \upsilon) \, I_0\left(\frac{\upsilon}{2} \right) + \upsilon \, I_1\left(\frac{\upsilon}{2} \right) \right] \,. \tag{6.39}$$

Using [149, eq. 13.3.2], [150, eq. A.1.19.c] we obtain

$$\begin{aligned} \int_0^\infty a^3 \exp\left(-\frac{a^2}{\lambda} \right) J_0\left(i \frac{2R}{\lambda_d} a \right) da &= \frac{\lambda^2 \Gamma(2)}{2\Gamma(1)} \, {}_1F_1(2; 1; \upsilon) \\ &= \frac{\lambda^2}{2} (1 + \upsilon) \, e^{\upsilon} \end{aligned} \tag{6.40}$$

Substituting (6.35)–(6.40) into (6.34) yields

$$\begin{aligned} r_{1j}(Y) &= \frac{b_{1j}}{\pi}\frac{\exp\left(-\frac{\gamma}{1+\xi}\right)}{1+\xi}\left\{G_j^2\,\gamma+\frac{\xi}{1+\xi}(1+\upsilon)-G_j\sqrt{\pi\,\upsilon}\exp\left(-\frac{\upsilon}{2}\right)\right. \\ &\qquad \left.\left[(1+\upsilon)\,I_0\left(\frac{\upsilon}{2}\right)+\upsilon\,I_1\left(\frac{\upsilon}{2}\right)\right]\right\} \\ &= \frac{b_{1j}}{\pi}\frac{\exp\left(-\frac{\gamma}{1+\xi}\right)}{1+\xi}\left\{G_j^2\,\gamma+\frac{\xi}{1+\xi}(1+\upsilon)-2\gamma\,G_j\,G_{STSA}\right\}. \end{aligned} \tag{6.41}$$

6.B Enhancement of Speech Signals Under Multiple Hypotheses Using an Indicator for Transient Noise Presence[3]

In this appendix, we formulate a speech enhancement problem under multiple hypotheses, assuming an indicator or detector for the transient noise presence is available in the short-time Fourier transform (STFT) domain. Hypothetical presence of speech or transient noise is considered in the observed spectral coefficients, and cost parameters control the trade-off between speech distortion and residual transient noise. An optimal estimator, which minimizes the mean-square error of the log-spectral amplitude, is derived, while taking into account the probability of erroneous detection. Experimental results demonstrate the improved performance in transient noise suppression, compared to using the optimally-modified log-spectral amplitude estimator.

6.B.1 Introduction

Enhancement of speech signals is of great interest in many voice communication systems, whenever the source signal is corrupted by noise. In a highly non-stationary noise environments, noise transients may be extremely annoying and significantly degrade the perceived quality and performances of subsequent coding or speech recognition systems. Existing speech enhancement algorithms, *e.g.*, [32, 33, 38], are generally inadequate for eliminating non-stationary noise components.

In some applications, an indicator for the transient noise activity may be available, *e.g.*, a siren noise in an emergency car, lens-motor noise of a digital video camera or a keyboard typing noise in a computer-based communication system. The transient spectral variances can be estimated in such cases from training signals. However, applying a standard estimator to the spectral coefficients may result in removal of critical speech components in case of falsely detecting the speech components, or under-suppression of transient noise in case of miss detecting the noise transients.

In this appendix, we formulate a speech enhancement problem under multiple hypothe-

[3]This appendix is based on [140].

ses, assuming some indicator or detector for the presence of noise transients in the STFT domain is available. Cost parameters control the trade-off between speech distortion and residual transient noise. We derive an optimal signal estimator that employs the available detector and show that the resulting estimator generalizes the optimally-modified log-spectral amplitude (OM-LSA) estimator [38]. Experimental results demonstrate the improved performance obtained by the proposed algorithm, compared to using the OM-LSA.

This appendix is organized as follows. In Section 6.B.2 we formulate the problem of spectral enhancement under multiple hypotheses. In Section 6.B.3 we derive the optimal estimator. In Section 6.B.4 we provide some experimental results and conclude in Section 6.B.5.

6.B.2 Problem formulation

Let $x(n)$, $d^s(n)$ and $d^t(n)$ denote speech and two uncorrelated additive interference signals, respectively, and let

$$y(n) = x(n) + d^s(n) + d^t(n) \tag{6.42}$$

be the observed signal. We assume that $d^s(n)$ is a quasi-stationary background noise while $d^t(n)$ is a highly non-stationary transient signal. The speech signal and the transient noise are not always present in the STFT domain, so we have four hypotheses for the noisy coefficients:

$$\begin{aligned}
H_{1s}^{\ell k} &: Y_{\ell k} = X_{\ell k} + D_{\ell k}^{s}\,, \\
H_{1t}^{\ell k} &: Y_{\ell k} = X_{\ell k} + D_{\ell k}^{s} + D_{\ell k}^{t}\,, \\
H_{0s}^{\ell k} &: Y_{\ell k} = D_{\ell k}^{s}\,, \\
H_{0t}^{\ell k} &: Y_{\ell k} = D_{\ell k}^{s} + D_{\ell k}^{t}\,,
\end{aligned} \tag{6.43}$$

where ℓ denotes the time frame index and k denotes the frequency-bin index.

In many speech enhancement applications, an indicator for the transient source may be available, *e.g.*, siren noise in an emergency car, keyboard typing in computer-based communication system and a lens-motor noise in a digital video camera. In such cases, *a*

priori information based on a training phase may yield a reliable detector for the transient noise. However, false detection of transient noise components when signal components are present may significantly degrade the speech quality and intelligibility. Furthermore, miss detection of transient noise components may result in a residual transient noise, which is perceptually annoying.

Let $\eta_j^{\ell k}$, $j \in \{0,1\}$ denote the detector decision in the time-frequency bin (ℓ, k), *i.e.*, a transient component is classified as a speech component under η_1 and as a noise component under η_0[4]. Let C_{10} denote the false-alarm cost with relation to the noise transient, *i.e.*, cost of making a decision η_0 when a noise transient is inactive or is not dominant w.r.t the speech component, and let the miss detection cost C_{01} be defined similarly. Let

$$d\left(x,y\right) \triangleq \left(\log|x| - \log|y|\right)^2 \tag{6.44}$$

denote the squared log-amplitude distortion function, let $A_{\ell k} \triangleq |X_{\ell k}|$ and let $R_{\ell k} \triangleq |Y_{\ell k}|$. Considering a realistic detector, we introduce the following criterion for the estimation of the speech expansion coefficient under the decision $\eta_j^{\ell k}$:

$$\begin{aligned}\hat{A}_{\ell k} = \arg\min_{\hat{A}} \Big\{ & C_{1j}\, p\left(H_{1s}^{\ell k} \cup H_{1t}^{\ell k} \,|\, \eta_j^{\ell k}, Y_{\ell k}\right) \\ & \times E\left[d\left(X_{\ell k}, \hat{A}\right) \,|\, Y_{\ell k}, H_{1s}^{\ell k} \cup H_{1t}^{\ell k}\right] \\ & + C_{0j}\, p\left(H_{0t}^{\ell k} \cup H_{0s}^{\ell k} \,|\, \eta_j^{\ell k}, Y_{\ell k}\right) d\left(G_{\min} R_{\ell k}, \hat{A}\right)\Big\}\end{aligned} \tag{6.45}$$

where the costs of perfect detection C_{00} and C_{11} are normalized to one. That is, under speech presence we aim at minimizing the MSE of the LSA. Otherwise, a constant attenuation $G_{\min} << 1$ is imposed for maintaining naturalness of the residual noise [38]. The cost parameters control the trade-off between speech distortion, consequent upon false detection of noise transients, and residual transient noise, resulting from miss detection of transient noise components.

6.B.3 Optimal estimation under a given detection

In this section we derive an optimal estimator for the speech signal under multiple hypotheses.

[4]Note that the detector is used for discriminating between transient speech components and transient noise components, and therefore not employed when transients are absent.

Spectral Estimation

We first reduce the problem into two basic hypotheses, $H_1^{\ell k}$ and $H_0^{\ell k}$. Under $H_1^{\ell k}$, the speech component is assumed present and more dominant than the noise component. This hypothesis includes $H_{1s}^{\ell k}$ as well as $H_{1t}^{\ell k}$ given that $|X_{\ell k}| \geq \beta\,|D_{\ell k}^t|$, where $\beta > 0$ is a predefined threshold parameter. The hypothesis $H_0^{\ell k}$ includes the cases $H_{0s}^{\ell k}$, $H_{0t}^{\ell k}$ and also $H_{1t}^{\ell k}$ with $|X_{\ell k}| < \beta\,|D_{\ell k}^t|$. Under $H_1^{\ell k}$ we estimate the speech in the MMSE-LSA sense, and under $H_0^{\ell k}$ we impose a constant attenuation to the noisy component. Note that ideally under $H_{1t}^{\ell k}$ an estimate for the speech component would be desired. However, if the noise transient is much more dominant we would better apply the constant low attenuation to the noisy component to avoid a strong residual noisy transient.

Let $p_{ij} \triangleq p\left(\eta_j^{\ell k} \,|\, H_i^{\ell k}\right)$. We are interested in detecting the interfering transient noise so p_{01} is the probability of a false alarm and p_{10} is the probability of miss detection. We assume that given any transient in the noisy coefficients, the detection error probability is independent of the observation and the signal-to-noise ratio (SNR). Therefore,

$$p\left(\eta_j^{\ell k} \,|\, H_i^{\ell k}, Y_{\ell k}\right) = p_{ij} \tag{6.46}$$

and

$$p\left(H_i^{\ell k} \,|\, \eta_j^{\ell k}, Y_{\ell k}\right) = p_{ij}\, p\left(H_i^{\ell k} \,|\, Y_{\ell k}\right) / p\left(\eta_j^{\ell k} \,|\, Y_{\ell k}\right)\ . \tag{6.47}$$

This assumption can be easily relaxed by employing a time-frequency dependent probability $p_{ij}^{\ell k}$. Considering the two basic hypotheses and substituting (6.47) into (6.45) we obtain

$$\begin{aligned}\hat{A}_{\ell k} = \arg\min_{\hat{A}} \Big\{ & p_{1j} C_{1j} p\left(H_1^{\ell k} \,|\, Y_{\ell k}\right) \\ & \times \int d\left(X_{\ell k}, \hat{A}\right) p\left(X_{\ell k} \,|\, Y_{\ell k}, H_1^{\ell k}\right) dX_{\ell k} \\ & +\ p_{0j} C_{0j} p\left(H_0^{\ell k} \,|\, Y_{\ell k}\right) d\left(G_{\min} R_{\ell k}, \hat{A}\right) \Big\}\ ,\end{aligned} \tag{6.48}$$

which yields

$$\begin{aligned}\log \hat{A}_{\ell k} \left[p_{1j} C_{1j} p\left(H_1^{\ell k} \,|\, Y_{\ell k}\right) + p_{0j} C_{0j} p\left(H_0^{\ell k} \,|\, Y_{\ell k}\right)\right] = \\ p_{1j} C_{1j} p\left(H_1^{\ell k} \,|\, Y_{\ell k}\right) E\left\{\log |X_{\ell k}| \,|\, Y_{\ell k}, H_1^{\ell k}\right\} \\ + p_{0j} C_{0j} p\left(H_0^{\ell k} \,|\, Y_{\ell k}\right) \log\left(G_{\min} R_{\ell k}\right)\ .\end{aligned} \tag{6.49}$$

Let $\xi_{\ell k}$ and $\gamma_{\ell k}$ denote the *a priori* and *a posteriori* SNRs, respectively[5], let $\upsilon_{\ell k} \triangleq \xi_{\ell k}\gamma_{\ell k}/\left(1+\xi_{\ell k}\right)$ and let

$$\begin{aligned}\Lambda\left(\xi_{\ell k},\gamma_{\ell k}\right) &\triangleq \frac{p\left(H_1^{\ell k}\right)}{p\left(H_0^{\ell k}\right)}\frac{p\left(Y_{\ell k}\,|\,H_1^{\ell k}\right)}{p\left(Y_{\ell k}\,|\,H_0^{\ell k}\right)} \\ &= \frac{p\left(H_1^{\ell k}\right)}{p\left(H_0^{\ell k}\right)}\frac{e^{\upsilon_{\ell k}}}{1+\xi_{\ell k}}\end{aligned} \tag{6.50}$$

denote the generalized likelihood ratio [33]. Accordingly,

$$p\left(H_1^{\ell k}\,|\,Y_{\ell k}\right) = \Lambda\left(\xi_{\ell k},\gamma_{\ell k}\right)/\left(1+\Lambda\left(\xi_{\ell k},\gamma_{\ell k}\right)\right)\,. \tag{6.51}$$

Let

$$\phi_j\left(\xi_{\ell k},\gamma_{\ell k}\right) = p_{1j}C_{1j}\Lambda\left(\xi_{\ell k},\gamma_{\ell k}\right) + p_{0j}C_{0j} \tag{6.52}$$

and let

$$G_{LSA}\left(\xi,\gamma\right) \triangleq \frac{\xi}{1+\xi}\exp\left(\frac{1}{2}\int_{\vartheta}^{\infty}\frac{e^{-t}}{t}dt\right) \tag{6.53}$$

denote the LSA gain function [34]. Then, combining the magnitude estimate $\hat{A}_{\ell k}$ with the phase of the noisy spectral coefficient $Y_{\ell k}$ we obtain an optimal estimate under the decision $\eta_j^{\ell k}$:

$$\begin{aligned}\hat{X}_{\ell k} &= \left[G_{\min}^{p_{0j}C_{0j}}G_{LSA}\left(\xi_{\ell k},\gamma_{\ell k}\right)^{p_{1j}C_{1j}\Lambda}\right]^{\phi_j^{-1}}Y_{\ell k} \\ &\triangleq G_{\eta_j}\left(\xi_{\ell k},\gamma_{\ell k}\right)\,Y_{\ell k}\,,\end{aligned} \tag{6.54}$$

where Λ and ϕ_j hold for $\Lambda\left(\xi_{\ell k},\gamma_{\ell k}\right)$ and $\phi_j\left(\xi_{\ell k},\gamma_{\ell k}\right)$, respectively.

In case of a decision η_1 (*i.e.*, transient component is classified as speech), the miss-detection cost C_{01} as well as the probabilities p_{01} and p_{11} control the trade-off between the attenuation associated with the hypothesis H_1 and the constant attenuation under speech absence, $G_{\min}$. Under a decision η_0, the trade-off is controlled by the false-alarm cost and the probabilities p_{00} and p_{10}.

Note that in case $p_{0j} = p_{1j}$ and $C_{0j} = C_{1j}$ for $j \in \{0,1\}$, the estimator (6.54) reduces to the OM-LSA estimator [38] under any of the detector decisions, since in that case the decision made by the detector does not contribute any statistical information.

[5]Note that the noise variance depends on whether a transient component is present or not. This will be specified in the next subsection.

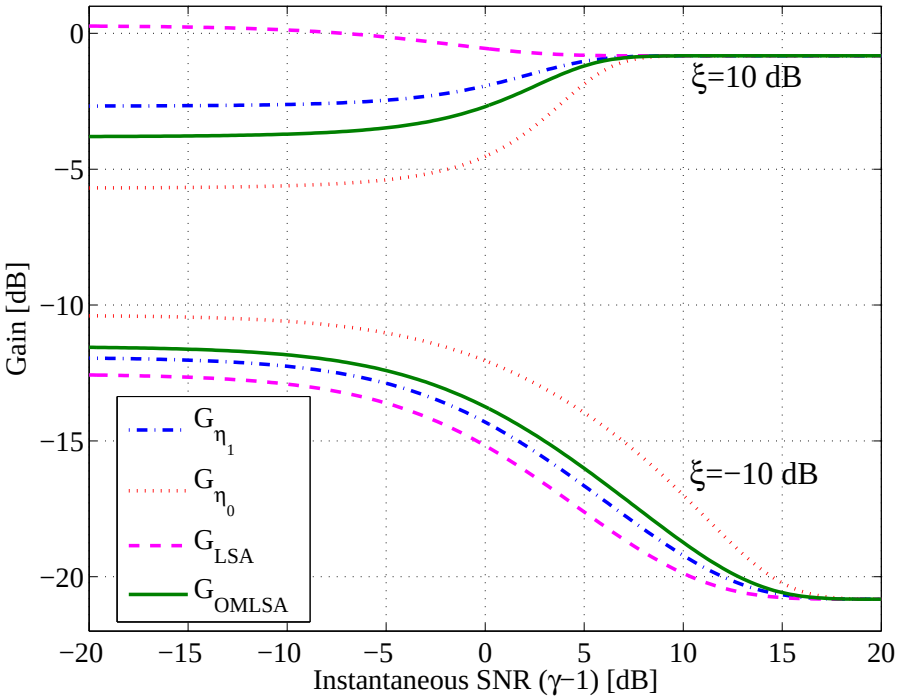


Figure 6.8: Gain curves for $p(H_1) = 0.8$, $C_{01} = 5$, $C_{10} = 3$, $G_{\min} = -15$ dB and false-detection and miss-detection probabilities of $p_{01} = p_{10} = 0.1$.

Figure 6.8 shows attenuation curves as a function of the *instantaneous* SNR, $\gamma - 1$, for different *a priori* SNRs. The detection-dependent gains G_{η_0} (dashed-dotted line) and G_{η_1} (dotted line) are compared to the LSA gain (dashed line) and the OM-LSA gain (solid line) [34, 38]. It shows that the cost parameters with the error probabilities of the detector shape the attenuation curve under any of the decisions made by the detector to compensate for any erroneous detection.

A priori and a posteriori SNR estimation

The spectrum of the background noise, $\lambda_{s,\ell k} \triangleq E\left\{|D^s_{\ell k}|^2\right\}$, can be estimated by using the minima-controlled recursive averaging algorithm [71]. The *a priori* signal-to-stationary noise ratio $\xi^s_{\ell k} \triangleq \lambda_{x,\ell k}/\lambda_{s,\ell k}$, where $\lambda_{x,\ell k} \triangleq E\left\{|X_{\ell k}|^2\right\}$, is practically estimated using the decision-directed approach [33, 38]. Given that a transient noise is present, the transient noise spectrum may be estimated from a training phase. Therefore, under η_0 we may estimate the *a priori* and *a posteriori* SNRs by using $\hat{\lambda}_{s,\ell k} + \hat{\lambda}_{t,\ell k}$ as the estimate for the noise spectrum [141], where $\lambda_{t,\ell k}$ is defined similarly to $\lambda_{s,\ell k}$. However, in case of an

erroneous detection, this approach may significantly distort the speech component, since both the *a priori* and *a posteriori* SNRs would be much smaller than their desired values. Therefore, we propose to smooth the noisy spectra

$$\zeta_{\ell k} = \mu\zeta_{\ell-1,k} + (1-\mu)\left|Y_{\ell k}\right|^2 , \tag{6.55}$$

with $0 < \mu < 1$. Accordingly, under a decision $\eta_0^{\ell k}$ we update the estimates such that

$$\begin{aligned} \eta_1^{\ell k} &: \hat{\xi}_{\ell k} = \hat{\xi}^s_{\ell k} , \quad \hat{\gamma}_{\ell k} = \hat{\gamma}^s_{\ell k} , \\ \eta_0^{\ell k} &: \hat{\xi}_{\ell k} = \hat{\xi}^s_{\ell k}\frac{\hat{\lambda}^s_{d,\ell k}}{\zeta_{\ell k}} , \quad \hat{\gamma}_{\ell k} = \hat{\gamma}^s_{\ell k}\frac{\hat{\lambda}^s_{d,\ell k}}{\zeta_{\ell k}} . \end{aligned} \tag{6.56}$$

As a result, the outcome of falsely detecting transient noise is less destructive since $\zeta_{\ell k}$ would be much smaller than $\lambda_{s,\ell k} + \lambda_{t,\ell k}$. However, in case of a perfect detection, $\zeta_{\ell k}$ is a reliable estimator for the noise spectrum given that μ is sufficiently small. In addition, under the existence of a high energy transient component we would like to further attenuate the noisy component to the level of the residual background noise. Therefore, under $\eta_0^{\ell k}$ we update $\tilde{G}_{\min} = G_{\min}\sqrt{\hat{\lambda}_{s,\ell k}/\zeta_{\ell k}}$.

6.B.4 Experimental results

In this section, we demonstrate the application of the proposed algorithm to speech enhancement in a computer-based communication system. The background office noise is slowly-varying while possible keyboard typing interference may exist. Since the keyboard signal is available to the computer, a reliable detector for the transient-like keyboard noise is assumed to be available based on a training phase but still, erroneous detections are reasonable. The speech signals are sampled at 16 kHz and degraded by a stationary background noise with 15 dB SNR and a keyboard typing noise such that the total SNR is 0.8 dB. The STFT is applied to the noisy signal with Hamming windows of 32 msec length and 75% overlap. The transient noise detector is assumed to have an error probability of 10% and the miss-detection and false-detection costs are set to 1.2. The weighting factor for the noisy spectra is $\mu = 0.5$.

Figure 6.9 demonstrates the spectrograms and waveforms of a signal enhanced by using the proposed algorithm, compared to using the OM-LSA algorithm. It can be seen that

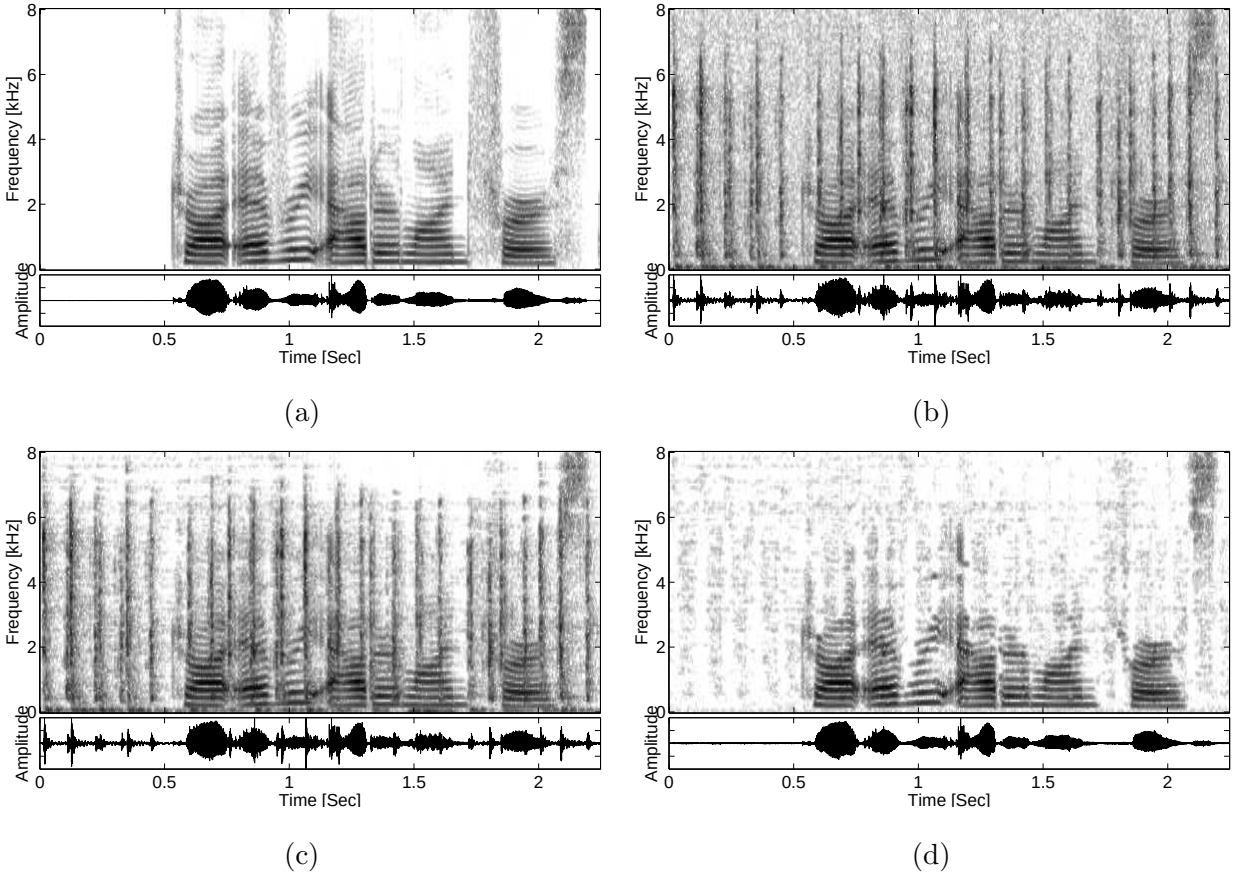


Figure 6.9: Speech spectrograms and waveforms. (a) Clean signal ("Draw any outer line first"); (b) noisy signal (office noise including keyboard typing noise, SNR=0.8 dB); (c) speech enhanced by using the OM-LSA estimator; (d) speech enhanced by using the proposed algorithm.

Table 6.3: Segmental SNR and Log Spectral Distortion Obtained Using the OM-LSA and the Proposed Algorithm.

Method	SegSNR [dB]	LSD [dB]	PESQ
Noisy speech	-2.23	7.69	1.07
OM-LSA	-1.31	6.77	0.97
Proposed Alg.	5.41	1.67	2.87

using our approach, the transient noise is significantly attenuated, while the OM-LSA is unable to eliminate the keyboard transients.

The objective evaluation includes three quality measures: segmental SNR (SegSNR), log-spectral distortion (LSD) and perceptual evaluation of speech quality (PESQ) score. The results are summarized in Table 6.3. It can be seen that the proposed detection and estimation approach significantly improves speech quality compared to using the OM-LSA algorithm. Informal listening tests confirm that the annoying keyboard typing noise is dramatically reduced and the speech quality is significantly improved.

6.B.5 Conclusions

We have introduced a new approach for a single-channel speech enhancement in a highly non-stationary noise environment where a reliable detector for interfering transients is available. The speech expansion coefficients are estimated under multiple-hypotheses in the MMSE-LSA sense while considering possible erroneous detection. The proposed algorithm generalizes the OM-LSA estimator and enables greater suppression of transient noise components.

Chapter 7

Single-Sensor Audio Source Separation Using Classification and Estimation Approach and GARCH Modeling[1]

In this chapter, we propose a new algorithm for single-sensor blind 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 classification and estimation approach, which enables to control the trade-off between residual interference and signal distortion. Experimental results demonstrate that for mixtures of speech and piano music signals, an improved source separation can be achieved compared to using Gaussian mixture model for both signals. The trade-off 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.

[1]This chapter is based on [151].

7.1 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 [85, 86], separation of different musical sources (e.g., different musical instruments) [85, 87, 88], separation of speech or singing voice from background music [89–92], and signal enhancement in nonstationary noise environments [35, 62, 93–95]. 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 [96] 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.

In [94, 95] 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 [87,88,90] the acoustic signals are modeled by Gaussian mixture models (GMMs), and in [35, 62] the acoustic signals are modeled by hidden Markov models (HMMs) with AR sub-sources. 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) [35, 88, 90], or by a maximum a posteriori (MAP) approach [62]. 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) [92, 152]. Recently, generalized autoregressive conditional heteroscedasticity (GARCH) models have been proposed for modeling speech signals for

speech enhancement [23, 24, 127, 133], speech recognition [26], and voice activity detection [27] 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 [1,5,25].

In this chapter, 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 trade-off 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 (non-optimal) 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 with several states of parameters enables smooth (diagonal) 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 trade-off between distortion of the desired signal caused by missed detection, and amount of residual interference resulting from false detection.

This chapter is organized as follows. In Section 7.2, 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 7.3, we define the GARCH codebook which is considered for speech signals and review the recursive conditional variance estimation. In Section 7.4, we describe the implementation of the proposed algorithm, and in Section 7.5 we provide some experimental results for audio separation of speech and music signals.

7.2 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.* [87–89] proposed a GMM for the signals' codebook in the short-time Fourier transform (STFT) domain, and in [35, 62, 94, 95] 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 $\{1, a_1, ..., a_p\}$ and innovation variance σ^2, the covariance matrix is $\sigma^2(A^T A)^{-1}$, where A is an $N \times N$ lower triangular Toeplitz matrix with $[1\, a_1 \,...\, a_p\, 0 \,...\, 0]^T$ as the first column. If the frame length N tends to infinity, the covariance matrix become circulant and hence diagonalized by the Fourier transform [35, 95]. 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 $p_1(i) \triangleq p(q_1 = i)$, $i = 1, ..., m_1$ and $p_2(j) \triangleq p(q_2 = j)$, $j = 1, ..., m_2$, 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 complex-valued Gaussian random vectors (see, e.g., [153, p. 89]) with known diagonal covariance matrices, i.e., $\mathbf{s}_1 \sim \mathcal{CN}\left(0, \Sigma_1^{(i)}\right)$ and $\mathbf{s}_2 \sim \mathcal{CN}\left(0, \Sigma_2^{(j)}\right)$.

Based on a given codebook, it is proposed in [88] and [95] to first find the active pair of states $\{i, j\} = \{q_1 = i, q_2 = j\}$ using a MAP criterion:

$$\left\{\hat{i}, \hat{j}\right\} = \arg\max_{i,j} p\left(\mathbf{x} \,|\, i, j\right) p\left(i, j\right) \tag{7.1}$$

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

$$\begin{aligned} \hat{\mathbf{s}}_1 &= E\left\{\mathbf{s}_1 \,|\, \mathbf{x}, \hat{i}, \hat{j}\right\} \\ &= \Sigma_1^{\left(\hat{i}\right)} \left(\Sigma_1^{\left(\hat{i}\right)} + \Sigma_2^{\left(\hat{j}\right)}\right)^{-1} \mathbf{x} \\ &\triangleq W_{\hat{i}\hat{j}}\, \mathbf{x} \end{aligned} \tag{7.2}$$

and similarly[2] $\hat{\mathbf{s}}_2 = W_{\hat{j}\hat{i}}\, \mathbf{x}$. Alternatively [35, 88, 90], the desired signal may be reconstructed in the mmse sense directly from

$$\begin{aligned} \hat{\mathbf{s}}_1 &= E\left\{\mathbf{s}_1 \,|\, \mathbf{x}\right\} \\ &= \sum_{i,j} p\left(i, j \,|\, \mathbf{x}\right) W_{ij}\, \mathbf{x}\,. \end{aligned} \tag{7.3}$$

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 $\mathbf{d} \sim \mathcal{CN}\left(0, \Sigma\right)$, 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 7.2.1 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

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

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 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 7.2.2, 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 (7.1). 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 *joint classification and estimation*. We show that in case of degenerated *simultaneous* classification and estimation formulation, closely related solutions can be derived under both approaches.

7.2.1 Simultaneous Classification and Estimation

Simultaneous detection and estimation formulation was first proposed by Middleton *et al.* [115, 116]. 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 [136]. 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 trade-off 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

$$C_{ij}^{\bar{i}\bar{j}}(\mathbf{s}, \hat{\mathbf{s}}) \triangleq b_{ij}^{\bar{i}\bar{j}} \|\mathbf{s} - \hat{\mathbf{s}}\|_2^2 \tag{7.4}$$

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

$$R = \sum_{i,j} \sum_{\bar{i},\bar{j}} \int \int C_{ij}^{\bar{i}\bar{j}}(\mathbf{s}_1, \hat{\mathbf{s}}_1)\, p(\mathbf{x} \,|\, \mathbf{s}_1, \bar{i}, \bar{j})\, p(\mathbf{s}_1 \,|\, \bar{i}, \bar{j})\, p(\bar{i}, \bar{j})\, p(\eta_{ij} \,|\, \mathbf{x})\, d\mathbf{s}_1\, d\mathbf{x}. \tag{7.5}$$

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

$$\min_{\eta_{ij}, \hat{\mathbf{s}}_1} \{R\}. \tag{7.6}$$

To derive a solution to (7.6) 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

$$r_{ij}^{\bar{i}\bar{j}}(\mathbf{x}, \hat{\mathbf{s}}_1) = \int C_{ij}^{\bar{i}\bar{j}}(\mathbf{s}_1, \hat{\mathbf{s}}_1)\, p(\mathbf{x} \,|\, \mathbf{s}_1, \bar{j})\, p(\mathbf{s}_1 \,|\, \bar{i})\, d\mathbf{s}_1 \tag{7.7}$$

denotes the average risk related to a decision η_{ij} when the true pair is $\{\bar{i}, \bar{j}\}$. The combined risk (7.5) can be written as

$$R = \sum_{i,j} \int p(\eta_{ij} \,|\, \mathbf{x}) \sum_{\bar{i},\bar{j}} p(\bar{i}, \bar{j})\, r_{ij}^{\bar{i}\bar{j}}(\mathbf{x}, \hat{\mathbf{s}}_1)\, d\mathbf{x}. \tag{7.8}$$

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

$$\hat{\mathbf{s}}_{1,ij} = \arg\min_{\hat{\mathbf{s}}_1} \sum_{\bar{i},\bar{j}} p(\bar{i}, \bar{j})\, r_{ij}^{\bar{i}\bar{j}}(\mathbf{x}, \hat{\mathbf{s}}_1). \tag{7.9}$$

Substituting (7.7) into (7.9) and setting the derivative to be equal to zero, we obtain the optimal estimate under η_{ij}:

$$\begin{aligned} \hat{\mathbf{s}}_{1,ij} &= \frac{\sum_{\bar{i}\bar{j}} b_{ij}^{\bar{i}\bar{j}}\, p(\mathbf{x} \,|\, \bar{i}, \bar{j})\, p(\bar{i}, \bar{j})\, W_{\bar{i}\bar{j}}\, \mathbf{x}}{\sum_{\bar{i}\bar{j}} b_{ij}^{\bar{i}\bar{j}}\, p(\mathbf{x} \,|\, \bar{i}, \bar{j})\, p(\bar{i}, \bar{j})} \\ &\triangleq G_{ij}\, \mathbf{x}. \end{aligned} \tag{7.10}$$

The derivation of (7.10) is given in Appendix 7.A. Note that in case the parameters $b_{ij}^{\bar{i}\bar{j}}$ are all equal, then the estimator (7.10) reduces to the mmse estimator (7.3) 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 (7.10) into (7.7), we obtain (see Appendix 7.B)

$$r_{ij}^{\bar{i}\bar{j}}(\mathbf{x}, \hat{\mathbf{s}}_1) = b_{ij}^{\bar{i}\bar{j}}\, p(\mathbf{x} \mid \bar{i}, \bar{j}) \left[\mathbf{x}^H \left(W_{\bar{i}\bar{j}}^2 - 2W_{\bar{i}\bar{j}} G_{ij} \right) \mathbf{x} + \mathbf{1}^T \Sigma_2^{(\bar{j})} W_{\bar{i}\bar{j}} \mathbf{1} \right] . \tag{7.11}$$

From (7.6) and (7.8), the optimal classification rule $\eta_{ij}(\mathbf{x})$ is obtained by minimizing the weighted average risks over all pairs of states:

$$\min_{ij} \sum_{\bar{i},\bar{j}} p(\bar{i}, \bar{j})\, r_{ij}^{\bar{i}\bar{j}}(\mathbf{x}, \hat{\mathbf{s}}_1) . \tag{7.12}$$

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 (7.12). 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 (7.10), and subsequently the optimal classifier chooses the appropriate pair (and the appropriate gain matrix) using (7.11) and (7.12). The combined solution guaranties minimum combined Bayes risk [116].

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 $q_1 \in \{0, 1, ..., m_1\}$ and $q_2 \in \{0, 1, ..., m_2\}$ 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

$$b_i^{\bar{i}} = \begin{cases} b_{1,\mathrm{m}} & i = 0,\ \bar{i} \neq 0 \\ b_{1,\mathrm{f}} & i \neq 0,\ \bar{i} = 0 \\ 1 & \text{o.w.} \end{cases} \tag{7.13}$$

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 (7.11)–(7.12), 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 (7.10), the high valued parameter raises the contribution of the corresponding state on the system estimate. If a parameter is smaller than one, than the chances of the detector to take this decision are higher, but, as the estimator (7.10) 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 distort 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 trade-off 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 (7.1)–(7.2), or the mmse estimator (7.3). However, the estimator (7.10) is optimal, not only when combined with the optimal classifier (7.12), but also when combined with any given classifier [116]. Therefore, this estimator may be combined with a sub-optimal classifier [e.g., the MAP classifier given by (7.1)] 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 non-ideal classifier and show that the same estimator (7.10) 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.

7.2.2 Joint Classification and Estimation

The application of a given classifier (e.g., a MAP classifier) followed by an estimator is shown in Figure 7.1. We denote this scheme as *joint classification and estimation*.

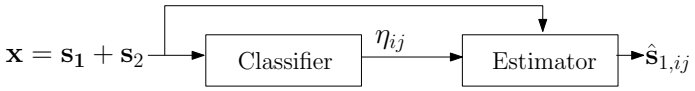


Figure 7.1: A cascade classification and estimation scheme.

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

$$\bar{\varepsilon_d^2}(\mathbf{x}) \triangleq p\left(q_1 = 1 \,|\, \mathbf{x}\right) E\left\{\|\mathbf{s}_1 - \hat{\mathbf{s}}_1\|_2^2 \,|\, q_1 = 1, \mathbf{x}\right\} \tag{7.14}$$

and the mean residual interference is defined by

$$\bar{\varepsilon_r^2}(\mathbf{x}) \triangleq p\left(q_1 = 0 \,|\, \mathbf{x}\right) E\left\{\|\mathbf{s}_1 - \hat{\mathbf{s}}_1\|_2^2 \,|\, q_1 = 0, \mathbf{x}\right\} . \tag{7.15}$$

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

$$\begin{aligned} \hat{\mathbf{s}}_{1,0j} &= \arg\min_{\hat{\mathbf{s}}_1} p\left(q_1 = 0 \,|\, \mathbf{x}\right) E\left\{\|\mathbf{s}_1 - \hat{\mathbf{s}}_1\|_2^2 \,|\, q_1 = 0, \mathbf{x}\right\} \\ \text{s.t.} \quad & \bar{\varepsilon_d^2}(\mathbf{x}) \leq \sigma_d^2 , \end{aligned} \tag{7.16}$$

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

$$\begin{aligned} \hat{\mathbf{s}}_{1,1j} &= \arg\min_{\hat{\mathbf{s}}_1} p\left(q_1 = 1 \,|\, \mathbf{x}\right) E\left\{\|\mathbf{s}_1 - \hat{\mathbf{s}}_1\|_2^2 \,|\, q_1 = 1, \mathbf{x}\right\} \\ \text{s.t.} \quad & \bar{\varepsilon_r^2}(\mathbf{x}) \leq \sigma_r^2 \end{aligned} \tag{7.17}$$

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 [57,104]. Under η_{0j} the Lagrangian is defined by (e.g., [154]):

$$L_d\left(\hat{\mathbf{s}}_1, \mu_d\right) = p\left(q_1 = 0 \,|\, \mathbf{x}\right) E\left\{\|\mathbf{s}_1 - \hat{\mathbf{s}}_1\|_2^2 \,|\, q_1 = 0, \mathbf{x}\right\} + \mu_d \left(\bar{\varepsilon_d^2}(\mathbf{x}) - \sigma_d^2\right) \tag{7.18}$$

and

$$\mu_d \left(\bar{\varepsilon}_d^2(\mathbf{x}) - \sigma_d^2\right) = 0 \quad \text{for} \quad \mu_d \geq 0\,. \tag{7.19}$$

Under η_{1j} the Lagrangian $L_r\left(\hat{\mathbf{s}}_1, \mu_r\right)$ 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., $L_d\left(\hat{\mathbf{s}}_1, \mu_d\right)$ or $L_r\left(\hat{\mathbf{s}}_1, \mu_r\right)$]. From $\bigtriangledown_{\hat{\mathbf{s}}_1} L_d\left(\hat{\mathbf{s}}_1, \mu_d\right) = 0$ we have[3]

$$\begin{aligned}
\hat{\mathbf{s}}_{1,0j} &= \frac{p\left(q_1 = 0 \,|\, \mathbf{x}\right) E\left\{\mathbf{s} \,|\, q_1 = 0, \mathbf{x}\right\} + \mu_d\, p\left(q_1 = 1 \,|\, \mathbf{x}\right) E\left\{\mathbf{s} \,|\, q_1 = 1, \mathbf{x}\right\}}{p\left(q_1 = 0 \,|\, \mathbf{x}\right) + \mu_d\, p\left(q_1 = 1 \,|\, \mathbf{x}\right)} \\
&= \frac{p\left(q_1 = 0 \,|\, \mathbf{x}\right) \sum_{\bar{j}} p\left(\bar{j} \,|\, q_1 = 0, \mathbf{x}\right) E\left\{\mathbf{s} \,|\, q_1 = 0, \bar{j}, \mathbf{x}\right\}}{\sum_{\bar{j}} p\left(q_1 = 0, \bar{j} \,|\, \mathbf{x}\right) + \mu_d \sum_{\bar{j}} p\left(q_1 = 1, \bar{j} \,|\, \mathbf{x}\right)} \\
&\quad + \frac{\mu_d\, p\left(q_1 = 1 \,|\, \mathbf{x}\right) \sum_{\bar{j}} p\left(\bar{j} \,|\, q_1 = 1, \mathbf{x}\right) E\left\{\mathbf{s} \,|\, q_1 = 1, \bar{j}, \mathbf{x}\right\}}{\sum_{\bar{j}} p\left(q_1 = 0, \bar{j} \,|\, \mathbf{x}\right) + \mu_d \sum_{\bar{j}} p\left(q_1 = 1, \bar{j} \,|\, \mathbf{x}\right)} \\
&= \frac{\sum_{\bar{j}} p\left(q_1 = 0, \bar{j} \,|\, \mathbf{x}\right) W_{0\bar{j}}\mathbf{x} + \mu_d \sum_{\bar{j}} p\left(q_1 = 1, \bar{j} \,|\, \mathbf{x}\right) W_{1\bar{j}}\mathbf{x}}{\sum_{\bar{j}} p\left(q_1 = 0, \bar{j} \,|\, \mathbf{x}\right) + \mu_d \sum_{\bar{j}} p\left(q_1 = 1, \bar{j} \,|\, \mathbf{x}\right)} \\
&= \frac{\sum_{\bar{i}\bar{j}} \tilde{\mu}_d^{\bar{i}}\, p\left(\bar{i}, \bar{j} \,|\, \mathbf{x}\right) W_{\bar{i}\bar{j}}\mathbf{x}}{\sum_{\bar{i}\bar{j}} \tilde{\mu}_d^{\bar{i}}\, p\left(\bar{i}, \bar{j} \,|\, \mathbf{x}\right)}
\end{aligned} \tag{7.20}$$

where

$$\tilde{\mu}_d^{\bar{i}} = \begin{cases} \mu_d & \bar{i} = 1 \\ 1 & \bar{i} = 0 \end{cases}\,. \tag{7.21}$$

Similarly, under signal-presence decision we have

$$\hat{\mathbf{s}}_{1,1j} = \frac{\sum_{\bar{i}\bar{j}} \tilde{\mu}_r^{\bar{i}}\, p\left(\bar{i}, \bar{j} \,|\, \mathbf{x}\right) W_{\bar{i}\bar{j}}\mathbf{x}}{\sum_{\bar{i}\bar{j}} \tilde{\mu}_r^{\bar{i}}\, p\left(\bar{i}, \bar{j} \,|\, \mathbf{x}\right)} \tag{7.22}$$

with

$$\tilde{\mu}_r^{\bar{i}} = \begin{cases} \mu_r & \bar{i} = 0 \\ 1 & \bar{i} = 1 \end{cases}\,. \tag{7.23}$$

Therefore, in general we can write

$$\hat{\mathbf{s}}_{1,ij} = \frac{\sum_{\bar{i}\bar{j}} \mu_i^{\bar{i}} p\left(\bar{i}, \bar{j} \,|\, \mathbf{x}\right) W_{\bar{i}\bar{j}}\mathbf{x}}{\sum_{\bar{i}\bar{j}} \mu_i^{\bar{i}} p\left(\bar{i}, \bar{j} \,|\, \mathbf{x}\right)} \tag{7.24}$$

with

$$\mu_i^{\bar{i}} = \begin{cases} \mu_d & i = 0,\, \bar{i} = 1 \\ \mu_r & i = 1,\, \bar{i} = 0 \\ 1 & \text{o.w.} \end{cases} \tag{7.25}$$

[3]Note that as shown in [57,104], there is no closed form solution for the value of the Lagrange multiplier. Instead it is used as a non-negative parameter.

and the estimator (7.24) is the same as (7.10) with $b_j^{\bar{j}} = 1$, $b_{1,\mathrm{m}} = \mu_d$, and $b_{1,\mathrm{f}} = \mu_r$ in (7.13). 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 (7.16) [or (7.17)] is inapplicable. Therefore, the problem reduces to a standard conditional mmse problem, which results in the estimator (7.2) which assumes a perfect classifier.

The main difference between the problem formulations in Sections 7.2.1 and 7.2.2 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.

7.3 GMM Vs. GARCH Codebook

In this section, we introduce a new codebook for mixtures of speech and music signals. GMM was used in [88–90] 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 [62,155] , or by using the k-means vector quantization algorithm [62,156]. Using a finite-state model with predetermined densities as in the case of GMM, mixture of AR models or HMM with AR sub-sources, 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 [92].

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) [1]. 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 [25]. GARCH model has been found useful for modeling speech signals in speech enhancement applications [24], [127, 133], speech recognition [26], and voice activity detection [27]. It has been shown [127] that

spectral variance estimation resulting from this model is a generalization of the decision-directed estimator [33] with improved tracking of the speech spectral volatility. Therefore, we propose in this work 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 $\mathbf{s}_2 \sim \mathcal{CN}\left(0, \Sigma_2^{(j)}\right)$. 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 $\Sigma_1^{(i_\ell)} = \text{diag}\left\{\boldsymbol{\lambda}_{\ell|\ell-1}^{(i_\ell)}\right\}$. 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:

$$\begin{aligned}\boldsymbol{\lambda}_{\ell|\ell-1}^{(i_\ell)} &= \lambda_{\min}^{(i_\ell)}\mathbf{1} + \alpha^{(i_\ell)}\mathbf{s}_1(\ell-1)\odot\mathbf{s}_1^*(\ell-1) \\ &\quad + \beta^{(i_\ell)}\left(\boldsymbol{\lambda}_{\ell-1|\ell-2}^{(i_{\ell-1})} - \lambda_{\min}^{(i_{\ell-1})}\mathbf{1}\right) \end{aligned} \tag{7.26}$$

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 $i_\ell = 0, 1, ..., m_1$ are sufficient conditions for the positivity of the conditional variance [127,133]. In addition, $\alpha^{(i_\ell)} + \beta^{(i_\ell)} < 1$ for all i_ℓ is a sufficient condition for a finite unconditional variance[4] [5]. The conditional density results from (7.26) 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 [7,8,127]. 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., $p(q_1(\ell) = i_\ell \,|\, q_1(\ell-1) = i_{\ell-1}) =$

[4]For necessary and sufficient condition see [121].

$p_1(i_\ell)$, similarly to the assumption used in [88] for the GMM approach.

It can be seen from (7.26) 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 (7.26). As we only have a mixed observation, we may estimate this random process of conditional variances based on the recursive estimation algorithm proposed in [127]. Assume that we have an estimate for the set of conditional variances at frame ℓ based on information up to frame $\ell-1$, $\hat{\Lambda}_\ell \triangleq \left\{\hat{\boldsymbol{\lambda}}_{\ell|\ell-1}^{(i_\ell)}\right\}_{i_\ell}$, then, following the model definition an mmse estimate of the next-frame conditional variance follows

$$\begin{aligned}\hat{\boldsymbol{\lambda}}_{\ell+1|\ell}^{(i_{\ell+1})} &= E\left\{\boldsymbol{\lambda}_{\ell+1|\ell}^{(q_1(\ell+1))} \mid q_1(\ell+1)=i_{\ell+1}, \hat{\Lambda}_\ell, \mathbf{x}(\ell)\right\} \\ &= \lambda_{\min}^{(i_{\ell+1})}\mathbf{1} + \alpha^{(i_{\ell+1})} E\left\{\mathbf{s}_1(\ell)\odot\mathbf{s}_1^*(\ell) \mid \hat{\Lambda}_\ell, \mathbf{x}(\ell)\right\} \\ &\quad + \beta^{(i_{\ell+1})}\left(E\left\{\boldsymbol{\lambda}_{\ell|\ell-1}^{(q_1(\ell))} \mid \hat{\Lambda}_\ell, \mathbf{x}(\ell)\right\} - E\left\{\lambda_{\min}^{(q_1(\ell))} \mid \hat{\Lambda}_\ell, \mathbf{x}(\ell)\right\}\mathbf{1}\right) \end{aligned} \quad (7.27)$$

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

$$\begin{aligned}\hat{\boldsymbol{\lambda}}_{\ell|\ell}^{(i_\ell,j_\ell)} &\triangleq E\left\{\mathbf{s}_1(\ell)\odot\mathbf{s}_1^*(\ell) \mid i_\ell, j_\ell, \hat{\Lambda}_\ell, \mathbf{x}(\ell)\right\} \\ &= \hat{\Sigma}_1^{(i_\ell)}\left(\hat{\Sigma}_1^{(i_\ell)}+\Sigma_2^{(j_\ell)}\right)^{-1}\left[\Sigma_2^{(j_\ell)}\mathbf{1}+\hat{\Sigma}_1^{(i_\ell)}\left(\hat{\Sigma}_1^{(i_\ell)}+\Sigma_2^{(j_\ell)}\right)^{-1}\mathbf{x}(\ell)\odot\mathbf{x}^*(\ell)\right] \end{aligned} \quad (7.28)$$

we obtain

$$\begin{aligned}E\left\{\mathbf{s}_1(\ell)\odot\mathbf{s}_1^*(\ell) \mid \hat{\Lambda}_\ell, \mathbf{x}(\ell)\right\} &= \sum_{i_\ell,j_\ell} p\left(i_\ell, j_\ell \mid \hat{\Lambda}_\ell, \mathbf{x}(\ell)\right) E\left\{\mathbf{s}_1(\ell)\odot\mathbf{s}_1^*(\ell) \mid i_\ell, j_\ell, \hat{\Lambda}_\ell, \mathbf{x}(\ell)\right\} \\ &= \sum_{i_\ell,j_\ell} p\left(i_\ell, j_\ell \mid \hat{\Lambda}_\ell, \mathbf{x}(\ell)\right)\hat{\boldsymbol{\lambda}}_{\ell|\ell}^{(i_\ell,j_\ell)} \end{aligned} \quad (7.29)$$

$$\begin{aligned}E\left\{\boldsymbol{\lambda}_{\ell|\ell-1}^{(q_1(\ell))} \mid \hat{\Lambda}_\ell, \mathbf{x}(\ell)\right\} &= \sum_{i_\ell,j_\ell} p\left(i_\ell, j_\ell \mid \hat{\Lambda}_\ell\right) E\left\{\boldsymbol{\lambda}_{\ell|\ell-1}^{(q_1(\ell))} \mid q_1(\ell)=i_\ell, \hat{\Lambda}_\ell, \mathbf{x}(\ell)\right\} \\ &\approx \sum_{i_\ell,j_\ell} p\left(i_\ell, j_\ell \mid \hat{\Lambda}_\ell\right)\hat{\boldsymbol{\lambda}}_{\ell|\ell-1}^{(i_\ell)} \end{aligned} \quad (7.30)$$

and

$$E\left\{\lambda_{\min}^{(q_1(\ell))} \mid \hat{\Lambda}_\ell, \mathbf{x}(\ell)\right\} = \sum_{i_\ell,j_\ell} p\left(i_\ell, j_\ell \mid \hat{\Lambda}_\ell\right)\lambda_{\min}^{(i_\ell)}. \quad (7.31)$$

A detailed recursive estimation algorithm is given in [127]. 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 [5, 7, 127] or may be evaluated as proposed in [133] such that each state would represent a different level of the optional dynamic range of the signals' 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 $\left\{\hat{\Sigma}_1^{(i_\ell)}\right\}_{i_\ell}$ 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.

7.4 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., [33, 34, 38]. The application of subband-based audio processing algorithms have been proposed for automatic speech recognition, e.g., [157, 158], speech enhancement [133], and also for single-channel source separation [152]. 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 [133].

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 $k \in \{1, ..., K\}$, $q_1 = 0$ 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 adjacent bands, as well as cross-band state probabilities, as proposed in [152].

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 [62, 156]. 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 [88,89,95]. For training the GARCH model we use the method proposed in [133]. 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, $i \in \{1, ..., m_1\}$, 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 [133] 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

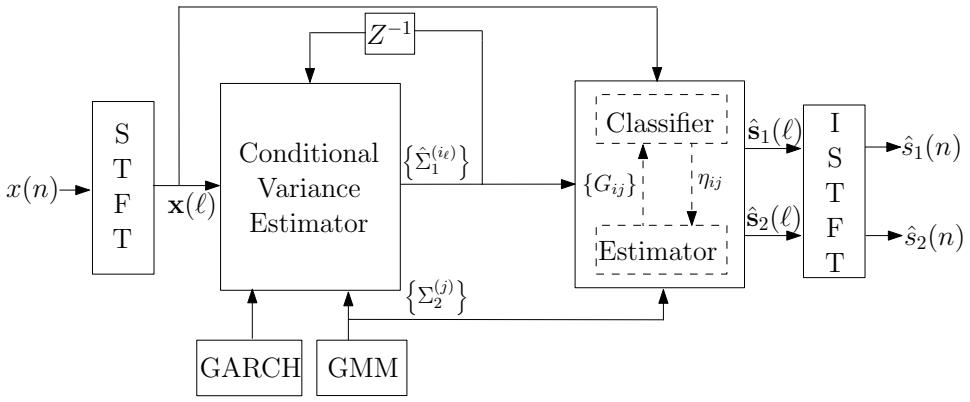


Figure 7.2: Block diagram of the proposed algorithm.

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 Figure 7.2 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 conditional covariance matrices $\left\{\hat{\Sigma}_1^{(i_\ell)} = \text{diag}\left(\hat{\boldsymbol{\lambda}}_{\ell|\ell-1}^{(i_\ell)}\right)\right\}_{i_\ell}$ are updated using (7.28) for any pair $\{i_\ell, j_\ell\}$, and then propagated one frame ahead using (7.27) 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 7.2.1, 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 7.2.2) may be applied as the classification and estimation block to enable a sub-optimal solution with lower computational cost. In fact, the computational complexity of this sub-optimal method is comparable to that of the mmse estimator (7.3) since the *a posteriori* probabilities required for the MAP classifier are used also in the estimation step.

7.5 Experimental Results

In this section we present experimental results for evaluating the performance of the proposed algorithm. In Section 7.5.1 we describe the experimental setup in our evaluation, and the objective quality measures. Then, in Section 7.5.2 we present experimental results. The experimental results are focused on (i) evaluating the performance of the proposed codebook compared to using a GMM-only model (while using mmse estimation for both codebooks), and (ii) evaluating the performance of the proposed simultaneous classification and estimation approach in the sense of signal distortion and residual interference.

7.5.1 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 [142] and include 8 different utterances by 8 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 seconds 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 sec 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 msec 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 7.4. 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 [88]. The

latter algorithm assumes a single-band GMM's 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. For each of the algorithms 4, 8 and 16 states are considered for the GMM, while the GARCH model is trained with up to 8 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 SNR (in the time domain) which is defined in dB by [143]

$$\text{SegSNR} = \frac{1}{|\mathcal{H}_1|} \sum_{\ell \in \mathcal{H}_1} \mathcal{T} \left\{ 10 \log_{10} \frac{\sum_{n=0}^{N-1} s^2 \left(n + \ell\, N/2\right)}{\sum_{n=0}^{N-1} \left[s \left(n + \ell\, N/2\right) - \hat{s} \left(n + \ell\, N/2\right)\right]^2} \right\}, \tag{7.32}$$

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 dB and 35 dB. The second quality measure is log-spectral distortion (LSD) which is defined in dB by [41]

$$\text{LSD} = \frac{1}{L} \sum_{\ell=0}^{L-1} \left\{ \frac{1}{N/2+1} \sum_{f=0}^{N/2} \left[10 \log_{10} \mathcal{C}\mathbf{s} \left(\ell, f\right) - 10 \log_{10} \mathcal{C}\hat{\mathbf{s}} \left(\ell, f\right)\right]^2 \right\}^{\frac{1}{2}}, \tag{7.33}$$

where $\mathbf{s}\left(\ell, f\right)$ denotes the fth element of the spectral vector $\mathbf{s}\left(\ell\right)$ (i.e., f denotes the frequency-bin index), $\mathcal{C}x \triangleq \max\left\{|x|^2, \epsilon\right\}$ is a spectral power clipped such that the log-spectrum dynamic range is confined to about 50 dB, that is, $\epsilon = 10^{-50/10} \times \max_{\ell,f} \left\{|\mathbf{s}(\ell, f)|^2\right\}$. Although the Segmental SNR and the LSD are common measures for speech enhancement, for the application of source separation it was proposed in [159, 160] to measure the signal to interference ratio (SIR). For source s_1 we may write

$$\hat{s}_1 = \zeta_1 s_1 + \zeta_2 s_2 + d\,. \tag{7.34}$$

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

$$\text{SegSIR} = \sum_{\ell} \mathcal{T} \left\{ 10 \log_{10} \frac{\zeta_1(\ell)^2 \sum_{n=0}^{N-1} s_1^2 \left(n + \ell\, N/2\right)}{\zeta_2(\ell)^2 \sum_{n=0}^{N-1} s_2^2 \left(n + \ell\, N/2\right)} \right\}, \tag{7.35}$$

where the parameters $\zeta_1(\ell)$ and $\zeta_2(\ell)$ are calculated for each segment as specified in [159].

The above mentioned measures attempt to evaluate the averaged performance of the algorithm. The proposed classification and estimation approach enables one to control the trade-off 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 trade-off 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 (7.33) and averaged only over $\mathcal{H}_1$. The interference reduction is evaluated in dB by [161]:

$$\mathrm{IR}_{H_0} = 10\log_{10}\frac{\sum_{\ell\in\mathcal{H}_0}\left\|\hat{\mathbf{s}}_1(\ell)\right\|^2}{\sum_{\ell\in\mathcal{H}_0}\left\|\mathbf{s}_2(\ell)\right\|^2}\,. \tag{7.36}$$

7.5.2 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 7.2.1, this also yields mmse estimation. Figure 7.3 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 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

[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.

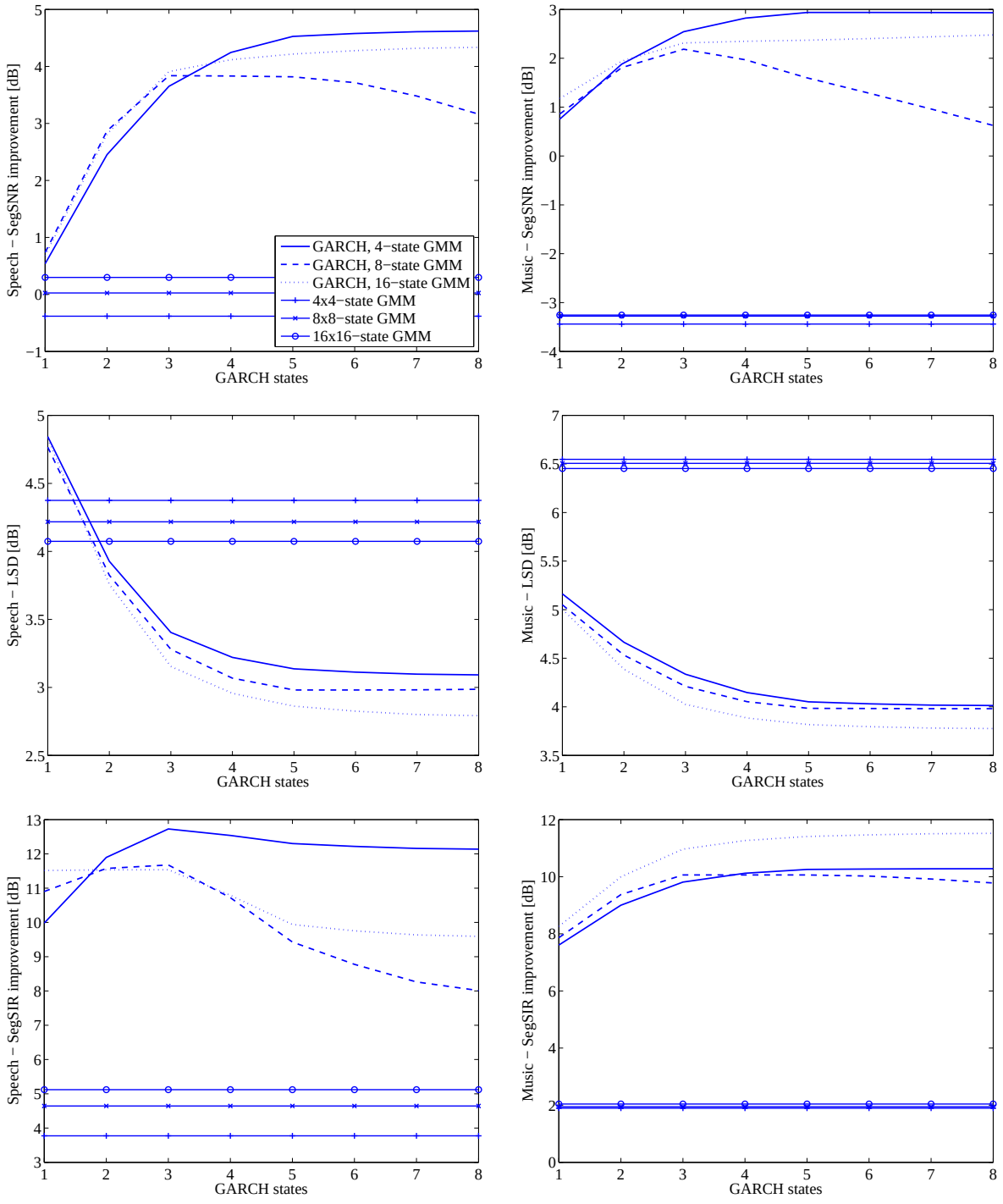
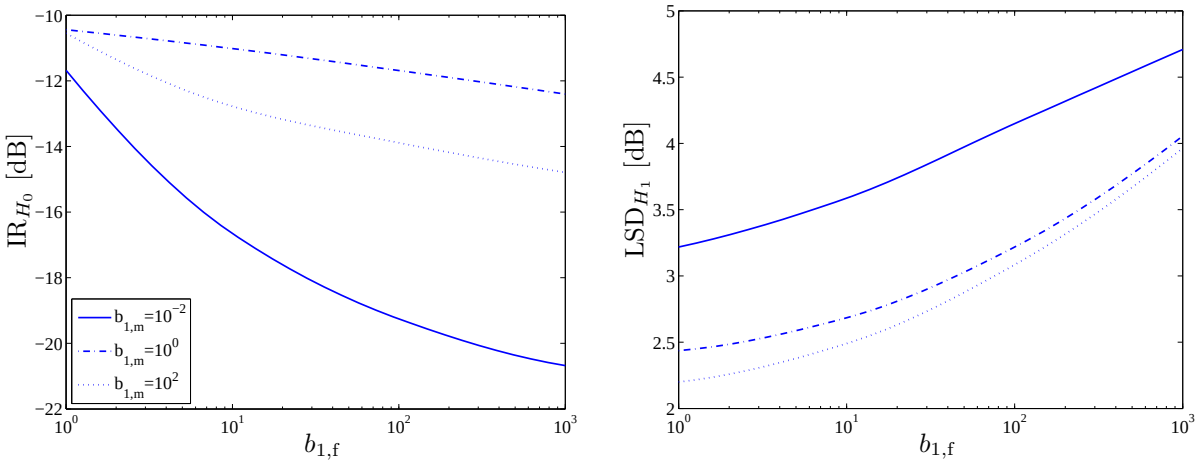


Figure 7.3: Quality measures for mmse estimation as functions of the number of GARCH states. The results (with different numbers of GMM states for the music signal) are compared with the GMM-based algorithm. Left column: results for speech signals; right column: results for music signals. Rows (from top to bottom): SegSNR improvement, LSD, and SegSIR improvement.



(a) (b)

Figure 7.4: Trade-off 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.

for some cases with 8-state GMM). However, the significant improvement is obtained by using up to 5 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.

Next, we verify the performance of the proposed simultaneous classification and estimation method. As this method enables one to control the trade-off 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. Figure 7.4 shows the trade-off 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 av-

Table 7.1: Averaged Quality Measures for the Estimated Speech Signals Using 3-state GARCH Model and 8-state GMM.

Parameters $[b_{1,\mathrm{m}}, b_{1,\mathrm{f}}, b_{2,\mathrm{m}}, b_{2,\mathrm{f}}]$	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 7.2: Averaged Quality Measures for the Estimated Music Signals Using 3-state GARCH Model and 8-state GMM.

Parameters $[b_{1,\mathrm{m}}, b_{1,\mathrm{f}}, b_{2,\mathrm{m}}, b_{2,\mathrm{f}}]$	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

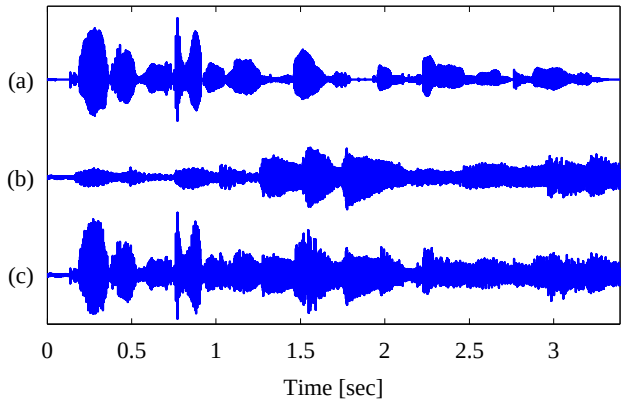


Figure 7.5: Original and mixed signals. (a) Speech signal: "Draw every outer line first, then fill in the interior"; (b) piano signal (*Für Elise*); (c) mixed signal.

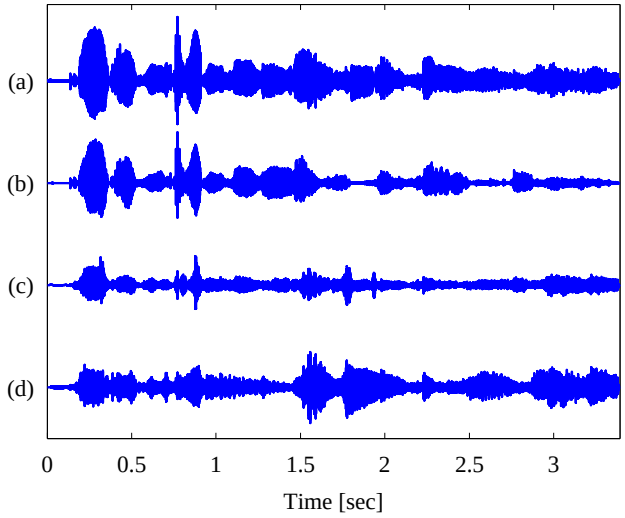


Figure 7.6: Separation of speech and music signals. (a) speech signal reconstructed by using the GMM-based algorithm (SegSNR improvement = 0.76, LSD = 3.77, SegSIR improvement = 1.29); (b) speech signal reconstructed using the proposed approach (SegSNR improvement = 2.46, LSD = 3.56, SegSIR improvement = 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).

eraged 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 3-state GARCH model and 8-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 trade-off between signal distortion and residual interference.

In Tables 7.1 and 7.2, we provide quality measures for both types of signals using different sets of parameters. This test was 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.

Figures 7.5 and 7.6 demonstrate the separation of a specific mixture of speech and piano signals. The speech waveform, the piano waveform and their mixture are shown in Figure 7.5, and Figure 7.6 shows the separated signals resulting from an 8-state GMM-based algorithm and from the proposed simultaneous classification and separation approach (using 3-state GARCH model for the speech signal, 8-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.

7.6 Conclusions

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 show that for mixture of speech and music (piano) signals, applying the proposed codebook significantly improves the separation results 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 trade-off between signal distortion and residual interference.

The proposed classification and estimation method may be advantageously utilized for other codebooks and 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.

7.A Derivation of (7.10)

By setting the derivative of $\sum_{\bar{i},\bar{j}} p\left(\bar{i},\bar{j}\right) r_{ij}^{\bar{i}\bar{j}}\left(\mathbf{x},\hat{\mathbf{s}}_1\right)$ in (7.9) to zero we obtain

$$\begin{aligned} 0 &= \sum_{\bar{i}\bar{j}} b_{ij}^{\bar{i}\bar{j}} p\left(\bar{i}\bar{j}\right) \Bigg[\hat{\mathbf{s}}_{1,ij} \int p\left(\mathbf{x} \,|\, \mathbf{s}_1, \bar{j}\right) p\left(\mathbf{s}_1 \,|\, \bar{i}\right) d\mathbf{s}_1 \\ &\quad - \int \mathbf{s}_1 \, p\left(\mathbf{x} \,|\, \mathbf{s}_1, \bar{j}\right) p\left(\mathbf{s}_1 \,|\, \bar{i}\right) d\mathbf{s}_1 \Bigg] \end{aligned} \tag{7.37}$$

where

$$\begin{aligned} p\left(\mathbf{x} \,|\, \mathbf{s}_1, \bar{j}\right) p\left(\mathbf{s}_1 \,|\, \bar{i}\right) &= p\left(\mathbf{x} \,|\, \mathbf{s}_1, \bar{i}, \bar{j}\right) p\left(\mathbf{s}_1 \,|\, \bar{i}, \bar{j}\right) \\ &= p\left(\mathbf{x} \,|\, \bar{i}, \bar{j}\right) p\left(\mathbf{s}_1 \,|\, \mathbf{x}, \bar{i}, \bar{j}\right) . \end{aligned} \tag{7.38}$$

Substituting (7.38) into (7.37) we obtain

$$\begin{aligned} 0 &= \sum_{\bar{i}\bar{j}} b_{ij}^{\bar{i}\bar{j}} p\left(\bar{i}\bar{j}\right) \left[\hat{\mathbf{s}}_{1,ij} p\left(\mathbf{x} \,|\, \bar{i}, \bar{j}\right)\right. \\ &\quad \left. - \; p\left(\mathbf{x} \,|\, \bar{i}, \bar{j}\right) E\left\{\mathbf{s}_1 \,|\, \mathbf{x}, \bar{i}, \bar{j}\right\}\right] \end{aligned} \tag{7.39}$$

and accordingly

$$\hat{\mathbf{s}}_{1,ij} = \frac{\sum_{\bar{i}\bar{j}} b_{ij}^{\bar{i}\bar{j}} \, p\left(\mathbf{x} \,|\, \bar{i}, \bar{j}\right) p\left(\bar{i}, \bar{j}\right) W_{\bar{i}\bar{j}} \, \mathbf{x}}{\sum_{\bar{i}\bar{j}} b_{ij}^{\bar{i}\bar{j}} \, p\left(\mathbf{x} \,|\, \bar{i}, \bar{j}\right) p\left(\bar{i}, \bar{j}\right)} \, . \tag{7.40}$$

7.B Derivation of (7.11)

The average risk is given by

$$\begin{aligned} r_{ij}^{\bar{i}\bar{j}}\left(\mathbf{x}, \hat{\mathbf{s}}_1\right) &= \int C_{ij}^{\bar{i}\bar{j}}\left(\mathbf{s}_1, \hat{\mathbf{s}}_1\right) p\left(\mathbf{x} \,|\, \mathbf{s}_1, \bar{j}\right) p\left(\mathbf{s}_1 \,|\, \bar{i}\right) d\mathbf{s}_1 \\ &= \int b_{ij}^{\bar{i}\bar{j}} \left\|\mathbf{s}_1 - \hat{\mathbf{s}}_1\right\|_2^2 \, p\left(\mathbf{x}, \mathbf{s}_1 \,|\, \bar{i}, \bar{j}\right) d\mathbf{s}_1 \, . \end{aligned} \tag{7.41}$$

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 $\Sigma_1 = \mathrm{diag}\{\sigma_1^2(1), \sigma_1^2(2), ..., \sigma_1^2(N)\}$ and $\Sigma_2 = \mathrm{diag}\{\sigma_2^2(1), \sigma_2^2(2), ..., \sigma_2^2(N)\}$. Following these notations we obtain

$$\begin{aligned} \int \|\mathbf{s}\|_2^2 \, p\left(\mathbf{x}, \mathbf{s}\right) d\mathbf{s} &= \sum_f \left\{ \int |\mathbf{s}(f)|^2 \, p\left(\mathbf{x}(f), \mathbf{s}(f)\right) d\mathbf{s}(f) \right. \\ &\quad \left. \times \prod_{f' \neq f} \int p\left(\mathbf{x}(f'), \mathbf{s}(f')\right) d\mathbf{s}(f') \right\} \end{aligned} \tag{7.42}$$

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 $\upsilon(f) \triangleq \xi(f)\gamma(f)/\left(1 + \xi(f)\right)$. By integrating over both the real and imaginary parts of $\mathbf{s}(f)$ and using [148, eq. 3.462.2] we obtain

$$\int |\mathbf{s}(f)|^2 \, p\left(\mathbf{x}(f), \mathbf{s}(f)\right) d\mathbf{s}(f) = \frac{\xi(f)\left(1 + \upsilon(f)\right)}{\pi \, \left(1 + \xi(f)\right)^2} \exp\left\{-\frac{\gamma(f)}{1 + \xi(f)}\right\} \tag{7.43}$$

and

$$\begin{aligned} \int p\left(\mathbf{x}(f'), \mathbf{s}(f')\right) d\mathbf{s}(f') &= p\left(\mathbf{x}(f')\right) \\ &= \frac{\exp\left\{-\frac{\gamma(f')}{1+\xi(f')}\right\}}{\pi \, \sigma_2^2(f') \left(1 + \xi(f')\right)} \, . \end{aligned} \tag{7.44}$$

Let $\Xi \triangleq \mathrm{diag}\{\xi(1), \xi(2), ..., \xi(N)\}$ and $V \triangleq \mathrm{diag}\{\upsilon(1), \upsilon(2), ..., \upsilon(N)\}$. Substituting (7.43) and (7.44) into (7.42) we obtain

$$\begin{aligned}
\int \|\mathbf{s}\|_2^2 \, p(\mathbf{x}, \mathbf{s}) \, d\mathbf{s} &= \sum_f \left\{ \frac{\xi(f)\,(1+\upsilon(f))}{\pi\,(1+\xi(f))^2} \exp\left(-\frac{\gamma(f)}{1+\xi(f)}\right) \right. \\
&\quad \left. \times \prod_{f' \neq f} \frac{1}{\pi\,\sigma_2^2(f')\,(1+\xi(f'))} \exp\left(-\frac{\gamma(f')}{1+\xi(f')}\right) \right\} \\
&= \frac{\mathbf{1}^T \Sigma_1 \,(I+V)\,(I+\Xi)^{-1}\,\mathbf{1}}{\pi^N \,|\Sigma_2\,(I+\Xi)|} \exp\left\{-\mathbf{x}^H \,(\Sigma_1+\Sigma_2)^{-1}\,\mathbf{x}\right\} . \quad (7.45)
\end{aligned}$$

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 [148, eq. 3.462.2] we obtain

$$\begin{aligned}
\int \left(\hat{\mathbf{s}}^H \mathbf{s} + \mathbf{s}^H \hat{\mathbf{s}}\right) p(\mathbf{x}, \mathbf{s}) \, d\mathbf{s} &= 2 \int \left(\hat{\mathbf{s}}_R^T \mathbf{s}_R + \mathbf{s}_I^T \hat{\mathbf{s}}_I\right) p(\mathbf{x}, \mathbf{s}) \, d\mathbf{s} \\
&= 2 \sum_f \left\{ g_{ij}(f) \int \left[\mathbf{x}_R(f)\mathbf{s}_R(f) + \mathbf{x}_I(f)\mathbf{s}_I(f)\right] p(\mathbf{x}(f), \mathbf{s}(f)) \, d\mathbf{s}(f) \right. \\
&\quad \left. \times \prod_{f' \neq f} \int p(\mathbf{x}(f'), \mathbf{s}(f')) \, d\mathbf{s}(f') \right\} \\
&= 2 \sum_f \frac{g_{ij}(f)\,\upsilon(f)\,\sigma_2^2(f)}{\pi} \prod_{f' \neq f} \frac{\exp\left\{-\frac{\gamma(f')}{1+\xi(f')}\right\}}{\pi\,\sigma_2^2(f')\,(1+\xi(f'))} \\
&= 2 \frac{\mathbf{1}^T G_{ij} \Sigma_2 V\,\mathbf{1}}{\pi^N \,|\Sigma_2\,(I+\Xi)|} \exp\left\{-\mathbf{x}^H \,(\Sigma_1+\Sigma_2)^{-1}\,\mathbf{x}\right\} . \quad (7.46)
\end{aligned}$$

Finally,

$$\begin{aligned}
\int p(\mathbf{x}, \mathbf{s}) \, d\mathbf{s} &= p(\mathbf{x}) \\
&= \frac{\exp\left\{-\mathbf{x}^H \,(\Sigma_1+\Sigma_2)^{-1}\,\mathbf{x}\right\}}{\pi^N \,|\Sigma_1+\Sigma_2|} . \quad (7.47)
\end{aligned}$$

Substituting (7.45)–(7.47) into (7.42) and using $W = \Xi\,(I+\Xi)^{-1}$, we obtain

$$\begin{aligned}
r_{ij}(\mathbf{x}, \hat{\mathbf{s}}_1) &= \bar{\bar{b_{ij}}}\, p(\mathbf{x}) \left[\mathbf{x}^H \left(W^2 - 2WG_{ij}\right) \mathbf{x} + \mathbf{1}^T \Sigma_2 W \mathbf{1}\right] \\
&= \frac{b_{ij}}{\pi^N \,|\Sigma_1+\Sigma_2|} \exp\left\{-\mathbf{x}^H \,(\Sigma_1+\Sigma_2)^{-1}\,\mathbf{x}\right\} \\
&\quad \times \left[\mathbf{x}^H \left(W^2 - 2WG_{ij}\right) \mathbf{x} + \mathbf{1}^T \Sigma_2 W \mathbf{1}\right] . \quad (7.48)
\end{aligned}$$

Chapter 8

Dual-Microphone Speech Dereverberation Using GARCH Modeling[1]

In this chapter, we develop a dual-microphone speech dereverberation algorithm for noisy environments, which is aimed at suppressing late reverberation and background noise. The spectral variance of the late reverberation is obtained with adaptively-estimated direct path compensation. A Markov-switching generalized autoregressive conditional heteroscedasticity (GARCH) model is used to estimate the spectral variance of the desired signal, which includes the direct sound and early reverberation. Experimental results demonstrate the advantage of the proposed algorithm compared to a decision-directed-based algorithm.

8.1 Introduction

In many speech communication systems the received signal is degraded by reverberation, as well as background noise. The reverberant signal consists of a direct sound, early reverberation, and late reverberation. Early reflections mainly contribute to coloration and tend to improve the intelligibility, whereas late reverberation causes a noise-like perception and degrades the fidelity and intelligibility of the speech signal.

[1]This chapter is based on [162].

Speech dereverberation algorithms can be divided into two classes. Algorithms in the first class are based on estimating and inverting the room impulse response (RIR), e.g., [163]. In the second class, algorithms try to suppress reverberation without estimating the RIR, e.g., [82]. Recently, Habets *et al.* [83] proposed a dual-microphone dereverberation system which is aimed at suppressing late reverberation that results from the tail of the RIR by applying a spectral enhancement approach. A direct path compensation (DPC) is applied to the late reverberant spectral variance estimate to enable better attenuation of the late reverberation with less distortion of the desired signal. However, the parameter of the DPC was evaluated directly from the RIR which is unknown in practice. In addition, the *a priori* signal to noise ratio (SNR) required for the spectral enhancement is estimated by using the traditional decision-directed approach. Recently, the generalized autoregressive conditional heteroscedasticity (GARCH) model with Markov regimes has been shown to be useful for speech enhancement applications [127, 133]. The model takes into account the strong correlation of successive spectral magnitudes, and is more appropriate than the decision-directed approach for speech spectral variance estimation in noisy environments.

In this chapter, we develop an improved dual-microphone speech dereverberation algorithm which relies on a Markov-switching GARCH (MS-GARCH) modeling of the desired early speech component, which consists of the direct sound and early reverberation. The model is applied to distinctive frequency subbands and specifies the volatility clustering of successive spectral coefficients, while a speech-absence state is used for evaluating the speech presence probability. Furthermore, an adaptive approach is developed to estimate the parameter for the DPC directly from the observed signals. Experimental results show that using the MS-GARCH modeling rather than the decision-directed approach, improved results can be obtained. Furthermore, by using the proposed algorithm, the performance obtained with blindly estimated DPC parameter is comparable to that obtained with an optimal DPC parameter that is calculated from the actual RIR, which is unknown in practice.

The chapter is organized as follows. In Section 8.2, we formulate the speech dereverberation problem and briefly review the algorithm proposed in [83]. In Section 8.3, we derive an adaptive estimator for the DPC parameter. In Section 8.4 we describe the

MS-GARCH model which is used for the desired signal, and in Section 8.5 we present some experimental results which demonstrate the improved performance of the proposed algorithm.

8.2 Dual-microphone dereverberation

Consider an $M-$microphone array located in a reverberant environment. Let $\mathbf{a}_m(n) = [a_{m,0}(n), ..., a_{m,L-1}(n)]^T$ denote the RIR at time n from the source signal $s(n)$ to the mth microphone, and let $d_m(n)$ denote the noise component received at the mth microphone. The observed signals are then given by

$$z_m(n) = \mathbf{a}_m^T(n)\,\mathbf{s}(n) + d_m(n) \tag{8.1}$$

where $\mathbf{s}(n) = [s(n), ..., s(n-L+1)]^T$. The RIR, $\mathbf{a}_m(n)$, can be divided into the direct path and early reflections, denoted by $\mathbf{a}_m^d(n)$, and late reflections, denoted by $\mathbf{a}_m^r(n)$. Accordingly,

$$a_{m,j}(n) = \begin{cases} a_{m,j}^d(n) & 0 \le j < t_r \\ a_{m,j}^r(n) & t_r \le j < L \end{cases}, \tag{8.2}$$

where t_r is the time where the late reverberation starts (about 40 to 80 ms). Hence, the reverberant signal can be divided into two signals

$$\mathbf{a}_m^T(n)\,\mathbf{s}(n) = x_m(n) + r_m(n)\ , \tag{8.3}$$

where $x_m(n)$ is the desired early speech component, and $r_m(n)$ denotes the late reverberant component. Applying the short-time Fourier transform (STFT) to the observed signals, we have

$$Z_m(\ell,k) = X_m(\ell,k) + R_m(\ell,k) + D_m(\ell,k)\ , \tag{8.4}$$

where ℓ represents the frame index, and k the frequency bin index. At the output of a delay and sum beamformer (DSB) which is steered towards the desired source, we have the time-frequency signal

$$Y(\ell,k) = X(\ell,k) + R(\ell,k) + D(\ell,k)\ . \tag{8.5}$$

Habets *et al.* [83] proposed a dual microphone dereverberation algorithm which is aimed at estimating the early speech component. In the system, shown in Figure 8.1, it is

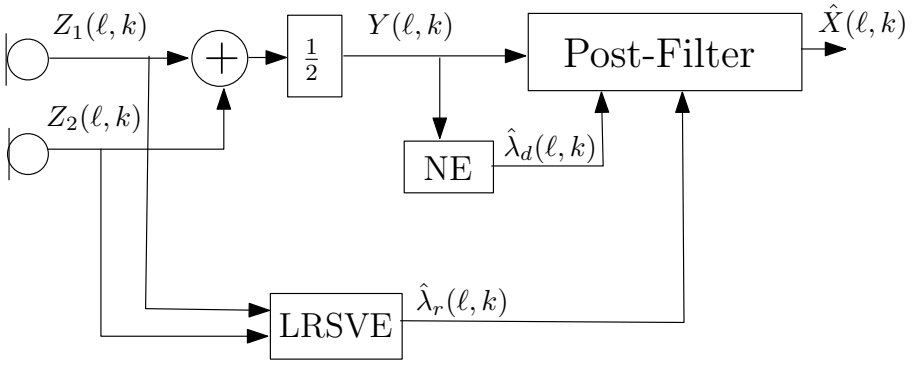


Figure 8.1: Dual microphone speech dereverberation system.

assumed that the arrival times of the direct speech signals are aligned. The lower branch is a late reverberant spectral variance estimator (LRSVE), $\hat{\lambda}_r(\ell,k)$, while the upper branch includes a beamformer, a background noise estimator (NE), $\hat{\lambda}_d(\ell,k)$, and a post-filter. The spectral variance of the noise signal, $\lambda_d(\ell,k)$, can be estimated, e.g., using [71]. The *a priori* SNR

$$\xi(\ell,k) = \frac{\lambda_x(\ell,k)}{\lambda_r(\ell,k) + \lambda_d(\ell,k)} \tag{8.6}$$

is estimated using the decision-directed approach [33].

The desired spectral coefficients are estimated by minimizing the mean square error of the log-spectral amplitude (LSA) [34] by assuming two hypotheses, speech presence (H_1) and absence (H_0). The resulting optimally-modified LSA estimator is given by [38]

$$\hat{X}(\ell,k) = G_{H_1}(\ell,k)^{p(\ell,k)}\, G_{H_0}(\ell,k)^{1-p(\ell,k)}\, Y(\ell,k)\ , \tag{8.7}$$

where $G_{H_1}(\ell,k)$ is the LSA gain under speech presence [34] and

$$G_{H_0}(\ell,k) = G_{\min}\frac{\hat{\lambda}_d(\ell,k)}{\hat{\lambda}_d(\ell,k) + \hat{\lambda}_r(\ell,k)} \tag{8.8}$$

to allow reduction of the late reverberant signal down to the noise floor [83]. In the next subsection we derive an adaptive estimator for the late reverberant spectral variance, and in Section 8.4 we formulate the MS-GARCH modeling applied for the desired signal. The speech presence probability $p(\ell,k)$ is discussed in Section 8.4.2.

8.3 Late reverberant spectral estimation

The spectral variance of the late reverberation at each microphone, $\lambda_{r,m}(\ell,k)$, can be obtained based on Polack's statistical reverberation model of the RIR [83], using an estimate of the spectral variance of the reverberant signal, $\lambda_{b,m}(\ell,k) = E\left\{|X_m(\ell,k)+R_m(\ell,k)|^2\right\}$. Let $T_{60}(k)$ denote the reverberation time of the room in the kth frequency band, let $\delta(k) = 3\ln(10)/T_{60}(k)$, let R denote the frame rate of the STFT, and let $\alpha(k) = \exp\{-2\delta(k)R/f_s\}$. Then, the spectral variance of the late reverberant signal $\lambda_r(\ell,k)$ at the output of the DSB is estimated by

$$\hat{\lambda}_r(\ell,k) = \frac{1}{2}\sum_{m=1}^{2}\alpha(k)^{\frac{t_r}{R}}\,\hat{\lambda}_{b,m}\left(\ell-\frac{t_r}{R},k\right). \tag{8.9}$$

However, to avoid over-estimation of $\lambda_r(\ell,k)$ when the source-microphone distance is smaller than the critical distance (i.e., the energy of the direct path is larger than the energy of all reflections) it was proposed to compensate the over estimation of the spectral variance of the reverberant signal using

$$\hat{\lambda}'_{b,m}(\ell) = \frac{\kappa_m(\ell)}{1+\kappa_m(\ell)}\alpha(k)\hat{\lambda}'_{b,m}(\ell-1,k) + \frac{1}{1+\kappa_m(\ell)}\hat{\lambda}_{b,m}(\ell,k)\,, \tag{8.10}$$

where $\kappa_m(\ell)$ is a compensation parameter which is related to the direct and reverberant energy at the mth microphone. The compensated estimate $\hat{\lambda}'_{b,m}(\ell)$ is then used in (8.9) as the spectral variance estimate of the reverberant signal. It was shown in [83] that applying this DPC prevents over-estimation of the late reverberant spectral variance and improves the quality of the output signal. However, the DPC parameter, κ_m, was calculated directly from the presumably known RIR. Here, we propose to estimate the parameter κ_m adaptively. In case κ_m is too large the spectral variance $\hat{\lambda}'_{b,m}(\ell,k)$ could become larger than $\hat{\lambda}_{b,m}(\ell,k)$, which indicates that over-estimation can occur and that the value of κ_m should be decreased. Furthermore, during the free-decay, which occurs after an offset of the source signal, $\hat{\lambda}'_{b,m}(\ell,k)$ should be equal to $\hat{\lambda}_{b,m}(\ell,k)$. Estimation of κ_m could therefore be performed after a speech offset. Unfortunately, the detection of speech offsets is rather difficult. However, we can conclude that κ_m should at least fulfill the following conditions: (i) $\hat{\lambda}_{b,m}(\ell,k) \geq \hat{\lambda}'_{b,m}(\ell,k)$, (ii) when speech is present and $\hat{\lambda}_{b,m}(\ell,k) < \hat{\lambda}'_{b,m}(\ell,k)$ the value of κ_m can be increased, (iii) when $\hat{\lambda}_{b,m}(\ell,k) > \hat{\lambda}'_{b,m}(\ell,k)$

the value of κ_m can be decreased slowly, and (iv) when $\hat{\lambda}_{b,m}(\ell,k) = \hat{\lambda}'_{b,m}(\ell,k)$ the value of κ_m is assumed to be correct. Therefore, we can update $\kappa_m(\ell)$ when speech is present using

$$\hat{\kappa}_m(\ell+1) = \max\left\{\hat{\kappa}_m(\ell) + \mu_\kappa\left(\frac{\sum_k \hat{\lambda}'_{b,m}(\ell,k)}{\sum_k \hat{\lambda}_{b,m}(\ell,k)} - 1\right),\, 0\right\}, \tag{8.11}$$

where μ_κ $(0 < \mu_\kappa < 1)$ denotes the step-size.

8.4 Modeling early reverberation using GARCH

Speech signals are characterized by time-varying energy levels and volatility. The spectral coefficients of the speech signal can be effectively characterized using an MS-GARCH model [127,133]. The GARCH parameters specify the volatility of the spectral coefficients, and the Markovian regimes allow the model to switch between different sets of GARCH parameters. Let $q_\ell \in \{0,...,Q\}$ denote the active state of a first-order Markov chain at frame ℓ with known state-transition probabilities. Let $\lambda_{x,q_\ell}(\ell,k\,|\,\ell-1)$ denote the conditional spectral variance of the desired signal $X(\ell,k)$ conditioned on q_ℓ and on all information up to previous frame, and let $\{V(\ell,k)\}$ be iid complex Gaussian random variables with zero-mean and unit variance. We assume that the spectral coefficients of the desired signal follow an MS-GARCH model [127], i.e., given q_ℓ

$$X(\ell,k) = \sqrt{\lambda_{x,q_\ell}(\ell,k\,|\,\ell-1)}\,V(\ell,k) \tag{8.12}$$

where

$$\begin{aligned}\lambda_{x,q_\ell}(\ell,k\,|\,\ell-1) &= \lambda_{\min,q_\ell} + \alpha_{q_\ell}\left|X(\ell-1,k)\right|^2 \\ &\quad + \beta_{q_\ell}\left[\lambda_{x,q_{\ell-1}}(\ell-1,k\,|\,\ell-2) - \lambda_{\min,q_{\ell-1}}\right]\end{aligned} \tag{8.13}$$

with $\lambda_{\min,q_\ell} > 0$ and α_{q_ℓ}, $\beta_{q_\ell} \geq 0$ for $q_\ell = 0,...,Q$. As can be seen from (8.12) and (8.13), the conditional spectral variances of successive frames at a specific frequency bin are strongly correlated. However, given the sequence of the conditional spectral variances and the active states, the spectral coefficients $\{X(\ell,k)\}$ are statistically independent. It was shown that the spectral variance estimation resulting from this model is a generalization

of the decision-directed estimator with improved tracking of the speech spectral volatility [127].

8.4.1 Spectral variance estimation

Let $\mathcal{Y}^\ell = \{Y(l,k) \mid l \leq \ell\}$ denote the set of the observed spectral coefficients up to frame ℓ. Given $\mathcal{Y}^\ell$ the set of conditional spectral variances can be recursively estimated using a propagation step

$$\begin{aligned}\hat{\lambda}_{x,q_\ell}(\ell,k \mid \ell-1) &= \lambda_{\min,q_\ell} + \alpha_{q_\ell} E\left\{|X(\ell-1,k)|^2 \mid \mathcal{Y}^{\ell-1}, q_\ell\right\} \\ &\quad + \beta_{q_\ell} E\left\{\lambda_x(\ell-1,k \mid \ell-2) \mid \mathcal{Y}^{\ell-1}, q_\ell\right\} \\ &\quad - \beta_{q_\ell} E\left\{\lambda_{\min,q_{\ell-1}} \mid \mathcal{Y}^{\ell-1}, q_\ell\right\} \end{aligned} \tag{8.14}$$

and an update step

$$\begin{aligned} E\left\{|X(\ell-1,k)|^2 \mid \mathcal{Y}^{\ell-1}, q_\ell\right\} &= \sum_{q_{\ell-1}} p\left(q_{\ell-1} \mid \mathcal{Y}^{\ell-1}, q_\ell\right) E\left\{|X(\ell-1,k)|^2 \mid \mathcal{Y}^{\ell-1}, q_{\ell-1}\right\} \\ &\triangleq \sum_{q_{\ell-1}} p\left(q_{\ell-1} \mid \mathcal{Y}^{\ell-1}, q_\ell\right) \hat{\lambda}_{x,q_{\ell-1}}(\ell-1,k \mid \ell-1)\,. \end{aligned} \tag{8.15}$$

A detailed estimation algorithm is given in [127]. The estimate of the spectral variance of the desired signal is then obtained by

$$\hat{\lambda}_x(\ell,k) = \sum_{q_\ell} p\left(q_\ell \mid \mathcal{Y}^\ell\right) \hat{\lambda}_{x,q_\ell}(\ell,k \mid \ell)\,. \tag{8.16}$$

Note that although the spectral variance is specified for each frequency bin independently, the Markovian state is frequency-independent. However, since different frequency bands of speech signals are characterized by different energy level and volatility, it was proposed in [133] to apply the model independently to distinctive subbands. Furthermore, a simple model estimation approach was proposed such that each state represents different energy level, and a specific state specifies signal absence. However, in our case the desired signal contains early reverberation such that the spectral variance at speech offsets has smoother decay than in case of a nonreverberant signal. Consequently, an immediate transition from a state which represents high spectral energy to a state which represents very low energy would not be expected. Therefore, the state transition probabilities are set such that the probability for a progressive state-transition is much higher than the probability for an immediate transition from the higher energy level to the lower.

8.4.2 Speech presence probability

The *posteriori* speech presence probability, $p(\ell,k)$, required for (8.7) is originally calculated [38] based on a Gaussian model from the *a priori* speech presence probability. The latter is evaluated based on the time-frequency distribution of the *a priori* SNR, $\xi(\ell,k)$. For a multi-sensor system, it was proposed in [83] to exploit the spatial information and to use additional parameter $P_{spatial}(\ell,k)$ for the *a priori* probability which is evaluated based on the spatial coherence between the microphone signals. In our case, the multi-state model for the speech spectral coefficients inherently results in a conditional probability for each state. Having a specific state for speech absence (say $q_\ell = 0$), we obtain a speech presence probability for each subband in each frame, $p\left(q_\ell \neq 0 \,|\, \mathcal{Y}^\ell\right)$. Accordingly, we define

$$P_{sb}(\ell,k) = \begin{cases} p_h & p\left(q_\ell \neq 0 \,|\, \mathcal{Y}^\ell\right) > T_h \\ p_l & p\left(q_\ell \neq 0 \,|\, \mathcal{Y}^\ell\right) < T_l \\ p\left(q_\ell \neq 0 \,|\, \mathcal{Y}^\ell\right) & \text{otherwise} \end{cases} \tag{8.17}$$

where $p_l \leq T_l \leq T_h \leq p_h$ are constrain parameters for the subband speech presence probability. The subband probability, $P_{sb}(\ell,k)$, is employed as an additional multiplicative parameter for the evaluation of the *a priori* speech presence probability. Note that although we do not use a specific index for the subband, $p\left(q_\ell \neq 0 \,|\, \mathcal{Y}^\ell\right)$ is calculated for each subband independently, and therefore $P_{sb}(\ell,k)$ includes also a frequency bin index.

8.5 Experimental results

In our experimental study, we consider synthetic RIRs which were generated using the *image* method. The speech signals, sampled at 8 kHz, include male and female speakers, each of 20 seconds. A moderate level of white Gaussian noise was added to each of the microphone signals. The distance between the two microphones is 0.15 meter, and the source-to-microphone distance was set to 0.5 and 1 meter (which are both smaller than the critical distance). While applying the MS-GARCH model, the model parameters are estimated from the noisy signal as proposed in [133].

Segmental signal to interference ratio (SegSIR) and log spectral distortion (LSD) are used to evaluate the performance of the proposed algorithm, as well as informal listening

Table 8.1: SegSIR and LSD obtained by using the decision-directed approach and the proposed MS-GARCH-based approach. $T_{60} = 0.5$ sec and d=0.5 meter. In parentheses - results using optimal DPC parameters.

	d=0.5 m, SNR=15 dB		d=0.5 m, SNR=20 dB	
	SegSIR [dB]	LSD [dB]	SegSIR [dB]	LSD [dB]
Unprocessed	5.849	4.875	7.284	2.681
Decision-directed	8.359	1.995	8.745	1.744
	(8.783)	(1.825)	(9.230)	(1.535)
MS-GARCH	9.010	1.700	9.392	1.493
	(9.265)	(1.606)	(9.715)	(1.367)

tests and inspection of spectrograms. For the quality measures, the direct sound signal was used as the reference signal. Figure 8.2 shows experimental results of the proposed algorithm as a function of the number of GARCH states, and for several reverberation times. The input SNR is 15 dB and the source to microphone distance is 0.5 m. It can be seen that the performance improves monotonically with the growth of the number of states, but, the most significant improvement is achieved by using up to 3 Markovian states.

Tables 8.1 and 8.2 compare the performance of the proposed algorithm with that of the original algorithm [83] which employs a decision-directed estimator for the *a priori* SNR. The reverberation time is $T_{60} = 0.5$ sec, and the proposed algorithm was applied with 3-state MS-GARCH model. In Table 8.1 the source to microphones distance is 0.5 meter and in Table 8.2 the distance is 1 meter. In both algorithms, the DPC parameters κ_1 and κ_2 are blindly estimated adaptively, as proposed in Section 8.3, and the results shown in parentheses are obtained using the optimal values which are evaluated from the actual RIRs. It can be seen that the GARCH modeling is more advantageous than the decision-directed approach, and the blindly estimated DPC parameters yield results which are comparable to using the optimal value.

In Figure 8.3 spectrogram and waveform of a noisy signal are shown with input SNR

Table 8.2: SegSIR and LSD obtained by using the decision-directed approach and the proposed MS-GARCH-based approach. T_{60} = 0.5 sec and d=1 meter. In parentheses - results using optimal DPC parameters.

	d=1 m, SNR=15 dB		d=1 m, SNR=20 dB	
	SegSIR [dB]	LSD [dB]	SegSIR [dB]	LSD [dB]
Unprocessed	2.295	6.379	2.864	4.578
Decision-directed	4.289	3.583	4.385	3.482
	(4.452)	(3.455)	(4.578)	(3.333)
MS-GARCH	4.551	3.521	4.654	3.442
	(4.941)	(3.390)	(5.110)	(3.298)

of 20 dB and a source to microphone distance of 1 m. The smearing caused by the late reverberation and the background noise are reduced.

Wave files are available online at: *http://siglab.technion.ac.il/~ari_a/Audio_demos.htm*.

8.6 Conclusions

We have developed a dual-microphone speech dereverberation algorithm for noisy environments which is based on MS-GARCH modeling of the desired early speech component. The spectral variance of the late reverberation is estimated from the observed signals while compensating for the energy of the direct path. The algorithm blindly operates in noisy and reverberant environments without any knowledge of the RIR, except for the reverberation time, which can be obtained blindly using, e.g., [164]. It is shown that compared to the original algorithm which employs the decision-directed estimator [83], improved performance is obtained with little distortion to the desired signal.

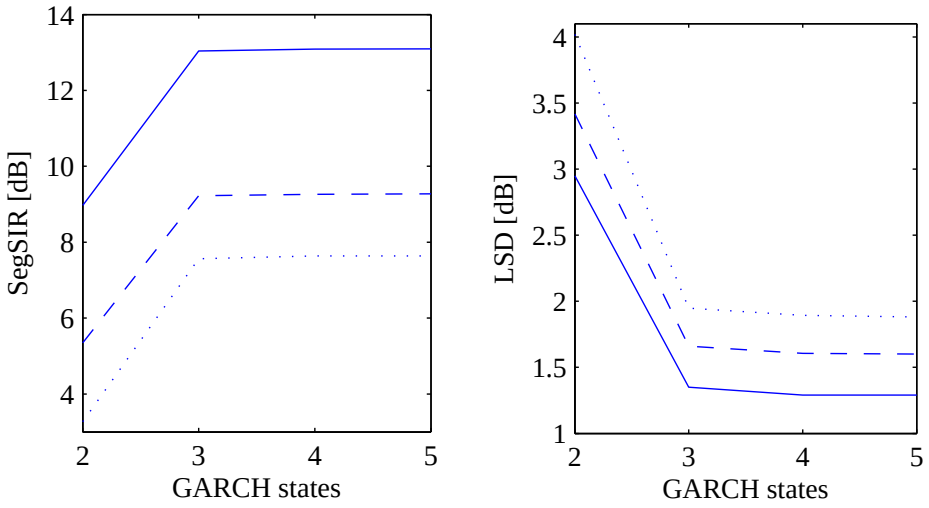
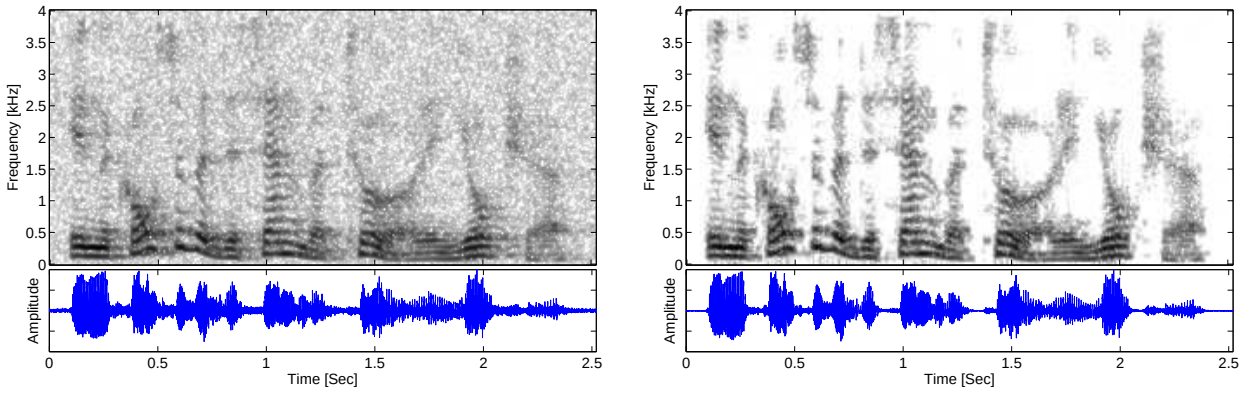


Figure 8.2: SegSIR and LSD as functions of the number of GARCH states (solid line: $T_{60} = 0.25$ sec, dashed line: $T_{60} = 0.5$ sec, and dotted line: $T_{60} = 0.75$ sec).



(a) (b)

Figure 8.3: Spectrograms and waveforms of (a) a noisy and reverberated speech signal, and (b) the processed signal.

Chapter 9

Research Summary and Future Directions

9.1 Research summary

In this thesis, we have introduced a new statistical model for nonstationary signals in the joint time-frequency domain, which is based on complex-valued GARCH model with Markov regimes. The model exploits the advantages of both the conditional heteroscedasticity structure of GARCH models and the time-varying characteristics of hidden Markov chains. We have developed conditions for finite second order moments and for asymptotic stationarity of the model, as well as for other MS-GARCH formulations which are used in econometrics. Moreover, we have developed recursive algorithms for the estimation of the conditional variance, as well as for signal restoration in noisy environment. A new formulation was proposed for the speech enhancement problem, based on simultaneous operations of speech detection and estimation. Considering the problem of a single-sensor audio source separation, we have generalized the simultaneous detection and estimation formulation to a multi-hypotheses case and incorporated the proposed MS-GARCH model for the speech signal. The result is a new algorithm for single-sensor audio source separation which is based on classification and estimation and GARCH modeling.

The main contributions of the thesis chapters are as follows:

In Chapter 3, we developed a comprehensive approach for stationarity analysis of MS-GARCH processes where finite-state-space Markov chains control the switching between

regimes, and GARCH models of order (p, q) are active in each regime. In case of processes with time-varying variances, conditions for asymptotic wide-sense stationarity are useful to ensure the existence of a finite asymptotic second-order moment. These conditions also show how some Markovian regimes can allow the conditional variance to grow over time and still the process will have a finite second-order-moment. Necessary and sufficient conditions for the asymptotic stationarity were obtained by constraining the spectral radius of representative matrices, which were built from the model parameters. These matrices also enabled derivation of compact expressions for the stationary variance of the processes.

Next, in Chapter 4, we have proposed a statistical model for nonstationary processes with time-varying volatility structure in the STFT domain such as speech signals. Exploiting the advantages of both the conditional heteroscedasticity structure of GARCH models and the time-varying characteristics of hidden Markov chains, we modeled the expansion coefficients as multivariate complex GARCH process with Markov-switching regimes. The correlation between successive coefficients in the time-frequency domain was taken into consideration by using the GARCH formulation which specifies the conditional variance as a linear function of its past values and past squared innovations. The time-varying structure of the conditional variance was determined by a hidden Markov chain which allows a different GARCH formulation in each state. We developed a recursive algorithm for estimating the signal and its conditional variance in the STFT domain from noisy observations. The conditional variance is recursively estimated for any regime by iterating propagation and update steps, while the evaluation of the regime conditional probabilities is based on the recursive correlation of the process. Experimental results demonstrated the improved performance of the proposed recursive algorithm compared to using an estimator which assumes a stationary process, even when the number of assumed regimes is smaller than the true number. When the number of assumed regimes approaches the true one, the recursive estimator yields comparable restoration results to those achievable by using the true model parameters. It was demonstrated that the recursive estimation approach has relatively small performance degradation compared to the theoretical estimation limit in the MMSE sense. Performance evaluation with real speech signals demonstrated better variance estimation when using a multi-regime model,

compared to using a single-regime model, and improved squared absolute value estimation in a noisy environment compared to using the decision-directed approach.

Chapter 5 addressed the problem of noncausal estimation. We developed state smoothing (i.e., noncausal state probability estimation) for MS-GARCH process, in which case the conditional variances depend on both past observations and the regime path. The state smoothing may be incorporated within the restoration algorithm to improve signal reconstruction as it employs further information. In addition, state smoothing may improve the probability evaluation for speech absence and therefore may result in improved VAD. Our noncausal state probability solution generalized both the standard forward-backward recursions and the stable backward recursion of HMP by capturing both the signal correlation along time and its conditioning on the regime path. Accordingly, we showed that the backward recursion requires two recursive steps for evaluating the conditional density of the given future observations corresponding to all optional future paths. Although the computational complexity of the generalized backward recursion grows exponentially with the delay, it was shown that a small number of future observations contribute with the most significant improvement to the state estimation.

In Chapter 6, a novel formulation of the single-channel speech enhancement problem was developed. The formulation relies on coupled operations of detection and estimation in the STFT domain, and a cost function that combines both the estimation and detection errors. A detector for the speech coefficients and a corresponding estimator for their values were jointly designed to minimize a combined Bayes risk. In addition, cost parameters enable to control the trade-off between speech quality, noise reduction and residual musical noise. The proposed method generalized the traditional spectral enhancement approach which considers estimation-only under signal presence uncertainty. In addition we have proposed a modified decision-directed *a priori* SNR estimator which is adapted to transient noise environment. Experimental results showed greater noise reduction with improved speech quality when compared with the STSA suppression rules under stationary noise. Furthermore, it was demonstrated that under transient noise environment, greater reduction of transient noise components may be achieved by exploiting reliable information for the *a priori* SNR estimation with simultaneous detection and estimation approach.

In Chapter 7, we have proposed a novel approach for a single-channel blind source separation of acoustic signals. The approach was based on classifying the mixed signal into appropriate sub-models related to a given codebook, and correspondingly estimate each of the sources. Unlike other classical methods which apply estimation alone, or distinctive operations of classification and estimation, in our method both operations were designed simultaneously, or the estimator was designed to allow compensation for erroneous classification. A new codebook was proposed for speech signals in the STFT domain based on the GARCH model. Accordingly, less restrictive pdf's are allowed in the STFT domain compared to GMM or AR-based model. Experimental results showed that for mixture of speech and music signals, applying the proposed codebook significantly improves the separation results 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 missed and false detection rates and the trade-off between signal distortion and residual interference.

Finally, in Chapter 8, we have developed a dual-microphone speech dereverberation algorithm for noisy environments, which was based on MS-GARCH modeling of the desired early speech component. The spectral variance of the late reverberation was estimated from the observed signals while compensating for the energy of the direct path. The algorithm blindly operates in noisy and reverberant environments without any knowledge of the RIR, except for the reverberation time. It was shown that compared with the original algorithm which employs the decision-directed estimator, improved performance was obtained with little distortion to the desired signal.

9.2 Future research directions

In this thesis, we have proposed a complex-valued MS-GARCH model and developed model-based algorithms for speech processing applications. Several directions may be interesting for future research. Here we discuss some of the main issues. More specific details are given in the conclusions of each chapter.

Multivariate GARCH model: The MS-GARCH model considered in this research formulates the correlation along time of successive spectral variances. In addition, all

frequencies in a specific subband share the same Markovian state and the same GARCH parameters. However, given their conditional variances, spectral coefficients at a specific frame are assumed statistically independent. A general formulation for a multivariate MS-GARCH may parameterizes statistical dependency between different frequencies at the same time-frame. Specifically, a single-state multivariate GARCH process $\boldsymbol{X}_t \in \mathbb{C}^N$ can be formulated as a zero-mean process with Λ_t covariance matrix which is given by [165,166]

$$\Lambda_t = C + \sum_{i=1}^{q} \left(\sum_{j=1}^{k} A_{ij} \boldsymbol{X}_{t-i} \boldsymbol{X}_{t-i}^H A_{ij}^H \right) + \sum_{i=1}^{p} \left(\sum_{j=1}^{k} B_{ij} \Lambda_{t-i} B_{ij}^H \right) . \tag{9.1}$$

To guarantee positive definiteness of Λ_t, C should be positive definite and A_{ij} and B_{ij} real valued matrices. This formulation allows non-diagonal covariance matrices such that different frequencies are correlated by the model definition. Considering voiced speech segments, modeling the correlation between different frequencies, such as between the pitch frequency and its harmonies, may significantly improve the performance of model-based algorithms.

Spectral variance estimation using speech detection: Integrating the simultaneous detection and estimation approach with MS-GARCH modeling for the spectral coefficients may improve both the conditional variance restoration and the detection operation. Specifically, a speech-absence state in the MS-GARCH formulation gives important information for speech presence. However, the spectral coefficients in some frequencies may be of negligible energy even under a speech-present state. Since the conditional variances are reconstructed recursively, the uncertainty assumption requires incorporation of a detection scheme within the propagation and update steps of the variance estimation to improve conditional variance estimation. Furthermore, since the MS-GARCH is a multi-state model, incorporation of a detection and estimation scheme requires generalization of the later approach to a multi-hypotheses case.

Multichannel speech processing: The proposed MS-GARCH modeling as well as the simultaneous detection and estimation approach may be employed for developing improved multichannel speech processing algorithms, such as multichannel speech enhancement, beamforming, and relative transfer function (RTF) identification.

A major drawback of many existing multichannel postfiltering techniques is that highly nonstationary noise components are not dealt with. The MS-GARCH model and the

simultaneous detection and estimation approach for the speech coefficients may be incorporated within the postfiltering of a beamformer. Considering a generalized sidelobe canceler scheme [167, 168], speech components are stronger at the beamformer output than in the noise reference signals, while noise components are strongest at the reference signals [107]. Accordingly, the beam-to-reference ratio may improve speech detection. In the blocking branch of the beamformer, which is aimed to create the noise reference signals, integrating a reliable detector may yield better reduction of both the coherent and incoherent noise at the beamformer output since the detection may improve the blocking of the speech components from leaking into the noise reference signals. While the detection and estimation which are applied in the postfiltering should be designed for better speech quality and perceptual intelligibility, the detection operation within the blocking branch should be designed for maximum blocking of speech components.

The proposed statistical model may also be useful for designing an RTF identification scheme that is adapted to speech signals. A detection and estimation scheme may be utilized to overcome the uncertainty of speech presence in the time-frequency domain. In time-frequency bins where speech components are detected, their PSD needs to be estimated as well as the RTF. However, under speech absence, only the cross-PSD of noise components may be estimated. Consequently, RTF identification performance as well as the rate of convergence may be improved.

Bibliography

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

[2] R. F. Engle, "Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation," *Econometrica*, vol. 50, no. 4, pp. 987–1007, July 1982.

[3] T. T. Baillie and T. Bollerslev, "Prediction in dynamic models with time-dependent conditional variances," *Journal of Econometrics*, vol. 52, pp. 91–113, 1992.

[4] J. D. Hamilton, "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometrica*, vol. 57, pp. 357–384, March 1989.

[5] ——, *Time Series Analysis*. Princeton University Press, 1994.

[6] S. F. Gray, "Modeling the conditional distribution of interest rates as a regime-switching process," *Journal of Financial Economics*, vol. 42, pp. 27–62, September 1996.

[7] F. Klaassen, "Improving GARCH volatility forecasts with regime-switching GARCH," *Empirical Economics*, vol. 27, no. 2, pp. 363–394, March 2002.

[8] M. Haas, S. Mittnik, and M. S. Paolella, "A new approach to Markov-switching GARCH models," *Journal of Financial Econometrics*, vol. 2, no. 4, pp. 493–530, Autumn 2004.

[9] J. Marcucci, "Forecasting stock Market volatility with regime-switching GARCH models," *Studies in Nonlinear Dynamics and Econometrics*, vol. 9, no. 4, Article 6, 2005.

[10] M. J. Dueker, "Markov switching in GARCH processes and mean reverting stock market volatility," *Journal of Business and Economic Statistics*, vol. 15, no. 1, pp. 26–34, January 1997.

[11] M. Frömmel, "Modelling exchange rate volatility in the run-up to EMU using Markov switching GARCH model," *Hannover University, Discussion paper*, no. 306, October 2004.

[12] J. Cai, "A Markov model of switching-regime ARCH," *Journal of Business and Economics Statistics*, vol. 12, no. 3, pp. 309–316, July 1994.

[13] J. D. Hamilton and R. Susmel, "Autoregressive conditional heteroskedasticity and changes in regime," *Journal of Econometrics*, vol. 64, pp. 307–333, July 1994.

[14] C. Francq and J.-M. Zakoïan, "The l^2-structure of standard and switching-regime garch models," *Stochastic Processes and their Applications*, vol. 115, no. 9, pp. 1557–1582, 2005.

[15] C. S. Wong and W. K. Li, "On a mixture autoregressive conditional heteroscedastic model," *Journal of the American Statistical Association*, vol. 96, pp. 982–985, September 2001.

[16] C. A. . E. Lazar, "Normal mixture GARCH(1,1) applications to exchange modelling," *ISMA Centre Discussion Papers in Finance 2004-06, The Business School for Financial Markets at University of Reading.*

[17] M. Haas, S. Mittnik, and M. S. Paolella, "Mixed normal conditional heteroskedasticity," *Journal of Financial Econometrics*, vol. 2, no. 2, pp. 211–250, 2004.

[18] M. Yang, "Some properties of vector autoregressive processes with Markov-switching coefficients," *Econometric Theory*, vol. 16, pp. 23–43, 2000.

[19] C. Francq and J.-M. Zakoïan, "Comments on the paper by Minxian Yang: "Some Properties of Vector Autoregressive Processes with Markov-Switching Coefficients"," *Econometric Theory*, vol. 18, pp. 815–818, 2002.

[20] ——, "Stationarity of multivariate Markov-switching ARMA models," *Journal of Econometrics*, vol. 102, pp. 339–364, 2001.

[21] J. Yao, "On squar-integrability of an AR process with Markov switching," *Statistics and Probability Letters*, vol. 52, pp. 265–27–, 2001.

[22] A. Timmermann, "Moments of Markov switching models," *Journal of Econometrics*, vol. 96, pp. 75–111, 2000.

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

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

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

[26] 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*, Pennsylvania ,Philadelphia, May 2005, pp. I.957–I.960.

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

[28] H. Drucker, "Speech processing in a high ambient noise environment," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, no. 2, pp. 165–168, June 1968.

[29] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process., ICASSP-79*, vol. 4, Apr. 1979, pp. 208–211.

[30] S. F. Boll, "Suppression of acousting noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[31] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[32] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.

[33] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[34] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

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

[36] ——, "Statistical-model-based speech enhancement systems," *Proceedings of The IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.

[37] V. I. Shin, D.-S. Kim, M. Y. Kim, and J. Kim, "Enhancement of noisy speech by using improved soft global decision," in *Eurospeech, Scandinavia*, 2001.

[38] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary environments," *Signal Processing*, vol. 81, pp. 2403–2418, Nov. 2001.

[39] W. Fong, S. J. Godsill, A. Doucet, and M. West, "Monte Carlo smoothing with application to audio signal enhancement," *IEEE Trans. Signal Processing*, vol. 50, no. 2, pp. 438–449, Feb. 2002.

[40] Y. Hu and P. C. Loizou, "A perceptually motivated approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 457–465, Sept. 2003.

[41] I. Cohen and S. Gannot, "Spectral enhancement methods," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Springer, 2007, ch. 45.

[42] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 5, pp. 497–514, Nov. 1997.

[43] Z. Goh, K. C. Tah, and T. G. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 3, pp. 287–292, May 1998.

[44] B. L. Lim, Y. C. Tong, and J. S. Chang, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 4, pp. 328–337, July 1998.

[45] H. Gustafsson, S. E. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and addptive averaging," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 8, pp. 799–807, Nov. 2001.

[46] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 6, pp. 341–351, Sept. 2002.

[47] R. Martin, "Speech enhancement based on minimum mean-square error estimation and super-Gaussian priors," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 845–856, Sept. 2005.

[48] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1218 – 1234, July 2006.

[49] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *Proc. 24th IEEE Int. Conf. Acoust., Speech Signal Process., ICASSP-99*, Phoenix, Arizona, Mar. 1999, pp. 789–792.

[50] Y. Ephraim and I. Cohen, "Recent advancements in speech enhancement," in *CRC Electrical Engineering Handbook*, 2005.

[51] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Proc. 23rd IEEE Int. Conf. Acoust., Speech Signal Process., ICASSP-98*, vol. 1, Seattle, Washington, May 1998, pp. 365–368.

[52] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[53] Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Processing Lett.*, vol. 8, no. 10, pp. 276–278, Oct. 2001.

[54] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 498–505, Sept. 2003.

[55] A. Davis, S. Nordholm, and R. Tongneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 412–423, Mar. 2006.

[56] J.-H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. Signal Processing*, vol. 54, no. 6, pp. 1965–1976, June 2006.

[57] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 4, pp. 251–266, July 1995.

[58] F. Asano, Hayamizu, T. Yamada, and Nakamura, "Speech enhancement based on the subspace method," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 5, pp. 497–507, Sept. 2000.

[59] F. Jabloun and B. Champagne, "A perceptual signal subspace approach for speech enhancement in colored noise," in *Proc. 27th IEEE Int. Conf. Acoust., Speech Signal Process., ICASSP-02*, Orlando, Florida, May 2002, pp. 569–572.

[60] H. Lev-Ari and Y. Ephraim, "Extension of the signal subspace speech enhancement approach to colored noise," *IEEE Signal Processing Lett.*, vol. 10, no. 4, pp. 104–106, Apr. 2003.

[61] B. H. Juang and L. R. Rabiner, "Mixture autoregressive hidden markov models for speech signals," *IEEE Trans. Speech Audio Processing*, vol. ASSP-33, no. 6, pp. 1404–1413, Dec. 1985.

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

[63] Y. Ephraim and W. J. Roberts, "Revisiting autoregressive hidden Markov modeling of speech signals," *IEEE Signal Processing Lett.*, vol. 12, no. 2, pp. 166–169, Feb. 2005.

[64] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *Bell Systems Technical Journal*, vol. 62, no. 4, pp. 1035–1074, Apr. 1983.

[65] N. Z. Tishby, "On the application of mixture ar hidden Markov models to text independent speaker recognition," *IEEE Trans. Signal Processing*, vol. 39, no. 3, pp. 563–570, Mar. 1991.

[66] Y. Ephraim, H. Lev-Ari, and W. J. J. Roberts, "A brief survey of speech enhancement," in *The Electronic Handbook*. CRC Press, 2005.

[67] R. Martin and C. Breithaupt, "Speech enhancement in the DFT domain using Laplacian speech priors," in *Proc. Int. Workshop on Acoust. Echo and Noise Control, IWAENC-03*, Kyoto, Japan, Sept. 2003, pp. 87–90.

[68] R. Martin, "Speech enhancement using mmse short time spectral estimation with Gamma distributed speech priors," in *Proc. 27th IEEE Int. Conf. Acoust., Speech Signal Process., ICASSP-02*, Orlando, Florida, May 2002, pp. I–253 – I–256.

[69] I. Cohen, "Relaxed statistical model for speech enhancement and a priori SNR estimation," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 870–881, Sept. 2005.

[70] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

[71] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 466–475, Sept. 2003.

[72] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 504–512, July 2001.

[73] E. A. P. Habets, *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement*. Ph.D. Thesis, Technische Universiteit Eindhoven, The Netherlands, June 2007.

[74] E. A. P. Habets and P. C. W. Sommen, "Speech dereverberation using spectral subtraction and a generalized statistical reverberation model," *Submitted to Elsevier's Speech Communication Journal*.

[75] E. A. P. Habets, S. Gannot, I. Cohen, and P. C. W. Sommen, "Joint dereverberation and residual echo suppression of speech signals in a noisy environment," *Submitted to IEEE Trans. on Audio, Speech, and Languages Processing*.

[76] P. Naylor and N. Gaubitch, "Speech dereverberation," in *Proc. Int. Workshop on Acoust. Echo and Noise Control, IWAENC-05*, Eindhoven, The Netherlands, Sept. 2005.

[77] S. Haykin, *Blind Deconvolution*, 4th ed. Prentice Hall information and system sciences series, 1994.

[78] M. Triki and D. T. M. Slock, "Delay and predict equalization for blind speech dereverberation," in *Proc. 31st IEEE Int. Conf. Acoust., Speech Signal Process., ICASSP-06*, Toulouse, France, May 2006, pp. 97–100.

[79] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise derevergeration using multichannel linear prediction," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 430–440, Feb. 2007.

[80] B. Radlovic, R. Williamson, and R. Kennedy, "Equalization in an acoustic reverberant environment: robustness results," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 311–319, May 2000.

[81] Y. Huang and J. Bnesty, "A class of frequency-domain adaptive approaches to blind multihannel identification," *IEEE Trans. Signal Processing*, vol. 51, no. 1, pp. 11–24, Jan. 2003.

[82] N. Gaubitch and P. Naylor, "Spatiotemporal averagingmethod for enhancement of reverberant speech," in *Proc. of the 15th International Conference on Digital Signal Processing (DSP 2007)*, July 2007, pp. 607–610.

[83] E. Habets, S. Gannot, and I. Cohen, "Dual-microphone speech dereverberation in a noisy environment," in *Proc. 6th IEEE Int. Symposium on Signal Process. and Information Technology, ISSPIT-2006*, Vancouver, Canada, Aug. 2006, pp. 651–655.

[84] K. Lebart and J. Boucher, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoustica*, pp. 359–366, 2001.

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

[86] 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, Hawaii, Apr. 2007, pp. 641–644.

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

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

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

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

[91] 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 on Acoust. Echo and Noise Control, IWAENC-06*, Paris, France, Sept. 2006, paper no. 52, pp. 1–4.

[92] 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, and Language Processing*, vol. 15, no. 5, pp. 1564–1578, July 2007.

[93] 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, Sept. 2004.

[94] 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*, vol. 1, Philadelphia, USA, Mar. 2005, pp. 1077–1080.

[95] ——, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 163–173, Jan. 2006.

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

[97] E. Vincent, "Musical source separation using time-frequency source priors," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 91–98, Jan. 2006.

[98] S. C. Douglas and X. Sun, "Convolutive blind separation of speech mixtures using the natural gradient," *Speech Communication*, vol. 39, pp. 65–78, 2003.

[99] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 320–327, May 2000.

[100] N. Mitianoudis and M. E. Davies, "Audio source separation in convolutive mixtures," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 489–497, Sept. 2003.

[101] G.-J. Jang and T.-W. Lee, "Single-channel signal separation using time-domain basis functions," *Journal of Machine Learning Research*, vol. 4, pp. 1365–1369, 2003.

[102] J. R. Hopgood and P. J. W. Rayner, "Single channel nonstationarity stochastic signal separation using time-varying filters," *IEEE Trans. Signal Processing*, vol. 51, no. 7, pp. 1739–1752, July 2003.

[103] K. V. Sørensen and S. V. Andersen, "Speech presence detection in the time-frequency domain using minimum statistics," in *Proc. of the 6th Nordic Signal Processing symposium - NORSIG*, Espoo, Finland, June 2004, pp. 340–343.

[104] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 4, pp. 334–341, July 2003.

[105] I. Potamitis and E. Fishler, "Microphone array voice activity detection and noise suppression using wideband generalized likelihood ratio," in *Eurospeech, Geneva*, Sept. 2003, pp. 525–528.

[106] J. Rosca, R. Balan, N. P. Fan, C. Beaugeant, and V. Gilg, "Multichannel voice detection in adverse environments," in *Proc. European Signal Process. Conf., EUSIPCO-02*, Toulouse, France, Sept. 2002, pp. 251–254.

[107] I. Cohen and B. Berdugo, "Multichannel signal detection based on transient beam-to-reference ratio," *IEEE Signal Processing Lett.*, vol. 10, no. 9, pp. 259–262, Sept. 2003.

[108] I. Potamitis, "Estimation of speech presence probability in the field of micfophone array," *IEEE Signal Processing Lett.*, vol. 11, no. 12, pp. 956–959, Dec. 2004.

[109] A. Spriet, M. Moonen, and J. Wouters, "The impact of speech detection errors on the noise reduction performance of multi-channel Wiener filtering and generalized sidelobe cancellation," *Signal Processing*, vol. 85, pp. 1073–1088, June 2005.

[110] T. G. Birdsall and J. O. Gobin, "Sufficient statistics and reproducing densities in simultaneous sequential detection and estimation," *IEEE Trans. Inform. Theory*, vol. IT-19, no. 6, pp. 760–768, Nov. 1973.

[111] J. Goutsias and J. M. Mendel, "Optimal simultaneous detection and estimation of filtered discrete semi-Markov chains," *IEEE Trans. Inform. Theory*, vol. 34, no. 3, pp. 551–568, May 1988.

[112] Z. Xie, C. K. Rushforth, R. T. Short, and T. K. Moon, "Joint signal detection and parameter estimation in multiuser communications," *IEEE Trans. Comput.*, vol. 41, no. 7, pp. 1208–1216, Aug. 1993.

[113] G. K. Kaleh and R. Vallet, "Joint parameter estimation and symbol detection for linear or nonlinear unknown channels," *IEEE Trans. Commun.*, vol. 42, no. 7, pp. 2406–2413, July 1994.

[114] B. Baygün and A. O. Hero III, "Optimal simultaneous detection and estimation under a false alarm constraint," *IEEE Trans. Inform. Theory*, vol. 41, no. 3, pp. 688–703, May 1995.

[115] D. Middleton and F. Esposito, "Simultaneous optimum detection and estimation of signals in noise," *IEEE Trans. Inform. Theory*, vol. IT-14, no. 3, pp. 434–444, May 1968.

[116] A. Fredriksen, D. Middleton, and D. Vandelinde, "Simultaneous signal detection and estimation under multiple hypotheses," *IEEE Trans. Inform. Theory*, vol. IT-18, no. 5, pp. 607–614, 1972.

[117] Y. Ephraim and N. Merhav, "Lower and upper bounds on the minimum mean-square error in composite source signal estimation," *IEEE Trans. Inform. Theory*, vol. 38, no. 6, pp. 1709–1724, Nov. 1992.

[118] R. W. Chang and J. C. Hancock, "On receiver structures for channels having memory," *IEEE Trans. Inform. Theory*, vol. IT-12, no. 4, pp. 463–468, Oct. 1966.

[119] G. Lindgren, "Markov regime models for mixed distributions and switching regressions," *Scan. J. Statist.*, vol. 5, pp. 81–91, 1978.

[120] M. Askar and H. Derin, "A recursive algorithm for the Bayes solution of the smoothing problem," *IEEE Trans. Automat. Contr.*, vol. AC-26, no. 2, pp. 558–561, Apr. 1981.

[121] 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.

[122] C. Francq, M. Roussignol, and J.-M. Zakoïan, "Conditional heteroskedasticity driven by hidden Markov chains," *Journal of Time Series Analysis*, vol. 22, no. 2, pp. 197–220, 2001.

[123] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[124] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd ed., W. Rheinbold, Ed. Academic Press, INC., 1985.

[125] L. A. Metzler, "A multiple-region theory of income and trade," *Econometrica*, vol. 18, no. 4, pp. 329–354, October 1950.

[126] D. Hawkins and H. Simon, "Note: some conditions of macroeconomics stability," *Econometrica*, vol. 17, no. 4, pp. 245–248, July 1949.

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

[128] ——, "Asymptotic stationarity of Markov-switching time-frequency GARCH processes," in *Proc. 31th IEEE Int. Conf. Acoust., Speech Signal Process., ICASSP-06*, Toulouse, France, May 2006, pp. III 452–445.

[129] Y. Ephraim and N. Merhav, "Hidden Markov processes," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1518–1569, June 2002.

[130] A. Abramson and I. Cohen, "State smoothing in Markov-switching time-frequency GARCH models," *IEEE Signal Processing Lett.*, vol. 13, no. 6, pp. 377–380, June 2006.

[131] P. Gill, W. Murray, and M. Wright, *Practical Optimization*. Academic Press, 1981.

[132] S. Han, "A globally convergent method for nonlinear programming," *Journal of Optimization Theory and Applications*, vol. 22, no. 4, pp. 297–309, 1977.

[133] A. Abramson and I. Cohen, "Markov-switching GARCH model and application to speech enhancement in subbands," in *Proc. Int. Workshop on Acoust. Echo and Noise Control, IWAENC-06*, Paris, France, Sept. 2006, paper no. 7, pp. 1–4.

[134] C. J. Kim, "Dynamic linear models with Markov-switching," *Journal of Econometrics*, vol. 60, pp. 1–22, 1994.

[135] H. Amiri, H. Amindavar, and R. L. Kirlin, "Array signal processing using GARCH noise modeling," in *Proc. 29th IEEE Int. Conf. Acoust., Speech Signal Process., ICASSP-04*, Montreal, Canada, Mar. 2004, pp. II–105–II–108.

[136] A. Abramson and I. Cohen, "Simultaneous detection and estimation approach for speech enhancement," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2348–2359, Nov. 2007.

[137] A. Subramanya, M. L. Seltzer, and A. Acero, "Automatic removal of typed keystrokes from speech signals," *IEEE Signal Processing Lett.*, vol. 14, no. 5, pp. 363–366, May 2007.

[138] W. A. Harrison, J. S. Lim, and E. Singer, "A new application of adaptive noise cancellation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 1, pp. 21–27, Feb. 1986.

[139] A. G. Jaffer and S. C. Gupta, "Coupled detection-estimation of gaussian processes in gaussian noise," *IEEE Trans. Inform. Theory*, vol. IT-18, no. 1, pp. 106–110, Jan. 1972.

[140] A. Abramson and I. Cohen, "Enhancement of speech signals under multiple hypotheses using an indicator for transient noise presence," in *Proc. 32nd IEEE Int. Conf. Acoust., Speech Signal Process., ICASSP-07*, Honolulu, Hawaii, Apr. 2007, pp. 553–556.

[141] E. Habets, I. Cohen, and S. Gannot, "MMSE log-spectral amplitude estimator for multiple interferences," in *Proc. Int. Workshop on Acoust. Echo and Noise Control., IWAENC-06*, Paris, France, Sept. 2006.

[142] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database," *Technical report, National Institute of Standards and Technology (NIST)*, Gaithersburg, Maryland (prototype as of December 1988).

[143] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Meaasures of Speech Quality.* Englewood Cliffs, NJ: Prentice-Hall, 1988.

[144] ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *International Telecommunication Union, Geneva, Switzerland*, Feb. 2001.

[145] A. Abramson Homepage. [Online]. Available: http:\\siglab.technion.ac.il\~ari_a

[146] A. Guérin, G. Faucon, and R. L. Bouquin-Jeannès, "Nonlinear acoustic echo cancellation based on voltera filters," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 6, pp. 672–683, Nov. 2003.

[147] E. Haensler and G. Schmidt, Eds., *Topics in Acoustic Echo and Noise Control*. Springer-Verlag, 2006.

[148] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 6th ed., A. Jefferey and D. Zwillinger, Eds. Academic Press, 2000.

[149] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, 2nd ed. Cambridge University Press, 1962.

[150] D. Middleton, *An Introduction to Statistical Comunication Theory*, 2nd ed. IEEE press, 1996.

[151] A. Abramson and I. Cohen, "Single-sensor blind source separation using classification and estimation approach and GARCH modeling," *submitted to IEEE Trans. Audio, Speech, and Language Processing*.

[152] 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, Canada, May 2004, pp. 641–644.

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

[154] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Athena Scientific, 1999.

[155] A. P. Dempster, N. M. Larid, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of The Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[156] G. A. F. Seber, *Multivariate Observations*. New York: Wiley, 1984.

[157] 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.

[158] 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.

[159] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[160] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte, "Proposals for performance measurement in source separation," in *Proc. 4th Int. Sym. on Independent Component Analysis and Blind Signal Separation (ICA)*, Nara, Japan, Apr. 2003, pp. 763–768.

[161] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[162] A. Abramson, E. A. P. Habets, S. Gannot, and I. Cohen, "Dual-microphone speech dereverberation using GARCH modeling," *to appear in Proc. 33th IEEE Int. Conf. Acoust., Speech Signal Process., ICASSP-08.*

[163] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Trans. Signal Processing*, vol. 51, no. 1, pp. 11–24, Jan. 2003.

[164] R. Ratnam, D. Jones, B. Wheeler, W. O'Brien Jr., C. Lansing, and A. Feng, "Blind estimation of reverberation time," *Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, Nov. 2003.

[165] R. F. Engle and K. F. Kroner, "Multivariate simultaneous generalized ARCH," *Econometrics Theory*, vol. 11, pp. 122–150, 1995.

[166] F. Comte and O. Lieberman, "Asymptotic theory for multivariate GARCH processes," *Journal of Multivariate Analysis*, vol. 84, no. 1, pp. 61–84, January 2003.

[167] S. Gannot and I. Cohen, "Adaptive beamforming and postfiltering," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Springer, 2007, ch. 48.

[168] I. Cohen, S. Gannot, and B. Berdugo, "An integrated real-time beamforming and post filtering system for nonstationary noise environments," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1064–1073, Oct. 2003.

מודלי GARCH בעלי משטר מיתוג מרקובי ויישומיהם לעיבוד ספרתי של אותות דיבור

ארי אברמסון

מודלי GARCH בעלי משטר מיתוג מרקובי ויישומיהם לעיבוד ספרתי של אותות דיבור

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר
דוקטור לפילוסופיה

ארי אברמסון

הוגש לסנט הטכניון - המכון טכנולוגי לישראל
טבת תשס״ח חיפה דצמבר 2007

המחקר נעשה בהנחיית פרופסור/ח' ישראל כהן
בפקולטה להנדסת חשמל

תודות

ברצוני להביע את תודתי והערכתי העמוקה לפרופ' ישראל כהן על הנחייתו המסורה. תודה על התמיכה המקצועית, על עידודך לשלמות ועל הרבה עצות מועילות לכל אורך שלבי המחקר.

תודה לשותפי למשרד קותי אברג'יל על הרבה דיונים פוריים. תודה לדר' עמנואל הבטס על דיונים מועילים ועל שאפשר לי להרחיב את המחקר לכיוון נוסף של ביטול הד אקוסטי.

תודה מיוחדת להורי מירי ומשה ולאהובתי אפרת על תמיכתם האינסופית לכל אורך הדרך.

אני מודה לטכניון, לקרן הלאומית למדע (מענק מס' 1085/05),ולאיחוד האירופי תחת פרוייקט Memories על התמיכה הכספית הנדיבה בהשתלמותי

תקציר

מחקר זה מתמקד בתיאוריה של מודלי GARCH (generalized autoregressive conditional heteroscedasticity) ויישומיהם לעיבוד אותות, ובמיוחד לעיבוד ספרתי של אותות דיבור.

מודל GARCH הוצג לראשונה על ידי Bollerslev ב-1986 כהכללה של מודל ARCH אותו פיתח Engle ב-1982. מודל זה מתאר באופן פרמטרי הפכפכות (שונות מותנית) התלויה בזמן, בעזרת ערכים קודמים של השונות המותנית ושל ריבועי ערכים קודמים של התהליך. המודל שימושי מאד בתחום מדעי הכלכלה עבור ניתוח וחיזוי מידת ההשתנות של תהליכים בשווקים פיננסיים. לאחרונה, מודל זה הוצע עבור יישומים לעיבוד אותות דיבור, כגון הדגשת אותות דיבור, זיהוי דובר, וכן גלאי לקיום דיבור. בעבודת מחקר זו, אנו מציעים ניסוח חדשני למודל GARCH מרוכב עם משטר מיתוג מרקובי (MS-GARCH) עבור אותות שאינם סטציונריים, במישור המשולב זמן-תדר. מודל זה מנצל הן את תכונות מודל GARCH המתאר תהליך בעל שונות מותנית המשתנה בזמן והן את מאפייני הדינאמיות הזמנית של שרשראות מרקוביות. המוטיבציה העיקרית למחקר זה מבוססת על תיאור סטטיסטי של אותות דיבור במישור זמן-תדר, כאשר דוגמאות ליישומים כוללות, לדוגמה, הפחתת רעש במערכות תקשורת (הן רעש רקע והן רעש רגעי - טרנזיינטי), הדגשת דיבור והפחתת הדהוד במערכות תקשורת בהן המיקרופון ממוקם הרחק מהדובר, וכן הפרדת מקורות קול הנקלטים במיקרופון יחיד.

התיזה מתחילה בניתוח סטציונריות אסימפטוטית של מודל MS-GARCH. תהליך MS-GARCH (כמו גם תהליך GARCH בעל מצב יחיד) אינו סטציונרי מכיוון שהמומנט מסדר שני משתנה בזמן. אולם, במידה ותהליכים אלו הינם סטציונריים אסימפטוטית במובן

הרחב, אזי מובטח כי השונות של התהליך תהיה בעלת גבול סופי. תנאים מספיקים והכרחיים לסטציונריות אסימפטוטית של תהליך GARCH פותחו על ידי Bollerslev, ואילו עבור מודלי MS-GARCH קיימים בספרות ניתוחי סטציונריות רק עבור ניסוחים מנוונים של המודל. גם עבור ניסוחים מנוונים אלו, בחלק מהמקרים פותחו תנאים שהינם הכרחיים אך לא בהכרח מספיקים לסטציונריות אסימפטוטית. במחקר זה, אנו מציעים ניתוח לסטציונריות של מודלי MS-GARCH המנוסחים באופן מלא ומפתחים תנאים שהינם הכרחיים ומספיקים לסטציונריות של מספר ניסוחי מודל כלליים, כפי שהוצעו בספרות.

מקדמי אותות דיבור במישור התמרת פורייה לזמן קצר מתאפיינים הן בחלקות של שונות המקדמים והן בפילוג בעל זנב עבה. שתי תכונות חשובות אלו מאפיינות גם תהליכי GARCH. תחת מוטיבציה זו, הוצע לאחרונה בתחום עיבוד אותות לנצל את מודל GARCH עבור מקדמי אותות דיבור במישור הזמן-תדר ליישומים של הדגשת דיבור. אולם, בעבודות אלו מניחים כי קיים גלאי למקדמי אות הדיבור ובנוסף מניחים כי פרמטרי המודל הינם קבועים בזמן. הנחה זו מגבילה את השימוש במודל במקרה של שינוי דובר המתאפיין בפרמטרים שונים, או אף בשינוי הברות. אנו מציעים ניסוח חדש למודל MS-GARCH עבור תהליכים אקראיים לא סטציונריים מרוכבים. מודל זה מתאים לייצוג אותות דיבור במישור התמרת פורייה לזמן קצר. כל מצב בשרשרת המרקובית החבויה מגדיר אוסף שונה של פרמטרים עבור מודל ה- GARCH ובכך מתאפשר ייצוג שונה עבור שינוי הברות ו/או דובר שעשויים להתאפיין בשינוי מצב השרשרת המרקובית. כאשר בוחנים תהליכים כלכליים, בדרך כלל אין בעיית רעש במדידת התהליך ולכן ערכי התהליך הנתונים עד לזמן כלשהו מספיקים בכדי לשחזר את השונות המותנית של התהליך באינדקס הזמן הבא. לעומת זאת, כאשר מנתחים אותות דיבור הנקלטים בסביבה רועשת, אין בידינו אלא מדידות רועשות של התהליך, ולפיכך לא ניתן לשחזר במדויק את ערכי התהליך ולא את השונות המותנית. לצורך כך פיתחנו אלגוריתם רקורסיבי לשערוך השונות המותנית של התהליך על סמך מדידות רועשות וכן אלגוריתם לשערוך האות עצמו. האלגוריתם מבוסס על שני צעדים, בדומה למסנן Kalman. בשלב הראשון, השונות המותנית מפועפעת צעד אחד קדימה בהתבסס על הגדרת המודל ואילו בשלב השני מתבצע עדכון השונות בעזרת המדידה הרועשת באותו הזמן. בנוסף, פיתחנו אלגוריתם לשערוך לא סיבתי של שרשרת המצבים המרקוביים מתוך מדידות

רועשות. האלגוריתמים הללו נמצאו יעילים עבור שחזור מקדמי אותות דיבור ליישומים של הדגשת אותות דיבור בסביבה רועשת וכן להפחתת הד אקוסטי. התוצאות שהתקבלו טובות יותר מאלו שהתקבלו בעזרת אלגוריתמים סטנדרטיים הקיימים בספרות.

תכונה חשובה נוספת של אותות דיבור הינה דלילות מקדמי הייצוג במישור התמרת פורייה לזמן קצר. פרט לעובדה שהדיבור אינו קיים בחלק מקטעי הזמן, גם בקטעי הזמן בהם הוא קיים, החלק הארי של האנרגיה נמצא בחלק קטן ממקדמי הייצוג. אלגוריתמים קיימים להדגשת אותות דיבור מניחים בדרך כלל משערך למקדמי הייצוג וכן גלאי נפרד לקיום המקדמים. במידה והגלאי מחליט כי מקדם ייצוג מכיל רכיב רעש בלבד, ניתן לדחות אותו במוצא מערכת השחזור. לחילופין, אלגוריתמים אחרים מבצעים שערוך המקדמים תחת חוסר ודאות. בגישה זו, מניחים הסתברות אפריורית לקיום מקדמי הייצוג של אות הדיבור ותחת הנחה זו מבצעים שערוך סטטיסטי של המקדמים. גישות של שילוב גלאי ומשערך מצומדים הפועלים יחדיו קיימות בספרות עבור עיבוד אותות והן עבור יישומים שונים בתקשורת. אנו מציעים בעבודת מחקר זו ניסוח חדשני לבעיית הדגשת אות דיבור. הניסוח כולל גלאי ומשערך הפועלים במקביל ואשר יחדיו ממזערים פונקצית מחיר מוכללת. שתי היפותזות מוצעות עבור מקדמי אותות הדיבור אשר מייצגות קיום או היעדרות של מקדמי הייצוג עבור כל מקדם במישור הזמן-תדר. פונקצית המחיר המוכללת כוללת מרכיב עבור שגיאת השערוך וכן מרכיב נוסף עבור שגיאת הגלאי. הגלאי והמשערך הינם מצומדים ומתוכננים יחדיו כדי למזער את תוחלת המחיר. על סמך ניסוח זה אנו מפתחים אלגוריתם להדגשת אותות דיבור תוך שימוש בפונקציות מחיר מתאימות. האלגוריתם המוצע הינו בעל ביצועים טובים יותר מאלגוריתם המבצע שערוך בלבד. יחד עם זאת, היעילות הגדולה בניסוח המוצע הינה היכולת לשלב גלאי חיצוני עבור רעש טרנזיינטי במערכת, תוך שמירת האופטימליות של המשערך גם בהינתן גלאי לא אידיאלי זה. אנו מציגים תוצאות המראות ששילוב האלגוריתם המוצע עם גלאי לא אידיאלי עבור רעש טרנזיינטי מאפשר הנחתה ניכרת של רעש זה תוך פגיעה קטנה בלבד ברכיבי אות הדיבור הרצוי.

בעיה מעניינת נוספת שנחקרה במסגרת תזה זו הינה הפרדת מקורות קול אשר נקלטו על ידי מיקרופון יחיד (דוברים שונים, כלים מוזיקליים שונים, או שילוב של שירה/דיבור עם מוזיקה). במידה ומספר מקורות נקלטים במספר מיקרופונים, ניתן לנצל את האינפורמציה ההדדית הקיימת באותות הנקלטים במיקרופונים השונים ו/או להפעיל גישות של סינון

מרחבי. בגישות אלו ניתן להגיע לעיתים לתוצאות משביעות רצון, גם כאשר מספר המקורות עולה על מספר המיקרופונים. אולם, כאשר מספר מקורות נקלטים במיקרופון יחיד יש צורך במידע אפריורי כלשהו על כל אחד מהמקורות כדי לאפשר הפרדה כלשהי. אלגוריתמים קיימים לפתרון בעיה זו מניחים מודל סטטיסטי שונה (מילון) עבור כל אחד מהאותות ומבצעים שערוך בגישה סטטיסטית של כל אחד מהאותות הרצויים. מילון זה נלמד בעזרת סדרת אימון המכילה אותות אשר אמורים לייצג מבחינה סטטיסטית את האות הרצוי. מודלים קיימים מבוססים על עירוב של פילוגים גאוסיים (GMM) או מודלים אוטו-רגרסיבים (AR) בעלי מספר סטים של מקדמי ייצוג. בכל אחד מהמקרים הללו, מטריצת השונות המשותפת האפשרית עבור וקטור נתון של הסיגנל מוגבלת לתת מרחב הנפרש על ידי הפילוגים המוגדרים במילון. עבור בעיה זו של הפרדת מקורות, אנו מציעים אלגוריתם חדש המבוסס על שימוש במודל GARCH והן על גישה המשלבת גילוי ושערוך. מאחר ועבור כל אחד מהאותות קיים מודל בעל מספר היפותזות, ההפרדה מבוצעת למעשה על ידי שילוב של מסווג ומשערך. שימוש בפרמטרי מחיר מאפשר למשתמש לשלוט על הפשרה בין עיוות האות, הנגרם כתוצאה מאי גילוי מקדמי האות הרצוי לבין הפרעה שיורית הנגרמת כתוצאה מגילוי שגוי. בעזרת אלגוריתם זה מתקבלת הפרדה טובה יותר מאשר בעזרת אלגוריתם קיים המניח מודל GMM בלבד והמבצע שערוך סטטיסטי בלבד.