Integration of Short-Time Fourier Domain Speech Enhancement and Observation Uncertainty Techniques for Robust Automatic Speech Recognition

Von der Fakultät IV Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Verleihung des akademischen Grades
Doktor-Ingenieur
genehmigte Dissertation

vorgelegt von Dipl.-Ing. Ramón Fernández Astudillo

Berlin 2010
D 83

Integration of Short-Time Fourier Domain Speech Enhancement and Observation Uncertainty Techniques for Robust Automatic Speech Recognition

Von der Fakultät IV Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Verleihung des akademischen Grades
Doktor-Ingenieur
genehmigte Dissertation

vorgelegt von Dipl.-Ing. Ramón Fernández Astudillo

Promotionausschuss:

Vorsitzender: Prof. Dr.-Ing. Sebastian Möller

Gutachter: Prof. Dr.-Ing. Reinhold Orglmeister

Gutachter: Prof. Dr.-Ing. Rainer Martin

Tag der wissenschaftlichen Aussprache: 1. Februar 2010
Berlin 2010
D 83

Acknowledgments

I would first like to thank my supervisor Prof. Orglmeister for his support and advice during the development of this thesis as well as Prof. Martin for accepting the co-supervision of my thesis and for his comments and corrections. I would also like to thank Prof. Möller for accepting to be the head of the evaluation comitee. Special thanks go to Dr. Kolossa for being an active research partner during this years and for her support and advice during the whole process.
This work would not have been possible without the additional founding of the "La Caixa" foundation and the German academic exchange service (DAAD). I am thankful for the trust they put on me. I´m also particularly thankful to Frau Paul-Walz from DAAD for her advice and support when dealing with the paperwork.
I would also like to thank all of the colleagues at the EMSP department who made my work easier by making it an excellent place to work. Special thanks go to Phillip Mandelartz for the developments achieved in his master thesis which were of great help for my dissertation.
Finally I would like to thank my father Luis, my mother Aurora, my brother Miguel and my girlfriend Marianna. This thesis is the result of a large process of learning and self understanding which I largely own to them. Thank you for being supportive all these years and particularly comprehensive with my lack of time.

Contents

1 Overview **1**

2 Speech Enhancement for Additive Noise and Gaussian a Priori Models **3**
2.1 Clean Speech Estimation . . . 5
2.1.1 Speech and Noise a Priori Models . . . 6
2.1.2 Likelihood Function . . . 7
2.1.3 Noisy Signal Distribution . . . 8
2.1.4 Wiener Filter Estimation of Amplitude and Phase . . . 9
2.1.5 Ephraim-Malah Amplitude Estimators . . . 11
2.1.6 Incorporation of Speech Presence Uncertainty in Speech Enhancement . . . 13
2.2 SNR Estimation . . . 14
2.2.1 Voice Activity Detection Based Noise Power Estimation . . . 14
2.2.2 Decision Directed a Priori Signal to Noise Ratio Estimation . . . 15
2.3 Derivations . . . 16
2.3.1 Complex Gaussian Distribution . . . 16
2.3.2 Rice Distribution . . . 17
2.3.3 Derivation of the Power Subtraction Estimator . . . 19
2.3.4 Derivation of the MMSE Criterion for Estimation . . . 20
2.3.5 Derivation of the Wiener Filter for Speech Enhancement . . . 21
2.3.6 Derivation of the First Ephraim-Malah Filter for Speech Enhancement . . . 22
2.3.7 Derivation of the Second Ephraim-Malah Filter for Speech Enhancement . . . 24

3 Automatic Speech Recognition with Hidden Markov Models **26**
3.1 Training of an ASR System . . . 28
3.2 Recognition . . . 28
3.3 Feature Extraction . . . 30
3.3.1 Mel-Cepstral Features . . . 31
3.3.2 Perceptual Linear Prediction Features . . . 32
3.3.3 RelAtive SpecTrAl Filtering . . . 34
3.3.4 Pre-Emphasis . . . 35
3.3.5 Delta and Acceleration Features . . . 35
3.3.6 Cepstral Mean Subtraction . . . 36
3.4 The ETSI ES 202 050 Standard for Robust Speech Recognition . . . 36
3.4.1 Log-Energy Features . . . 38

3.4.2 Blind Equalization . . . 38
3.5 Measuring Recognition Performance . . . 39
3.6 Automatic Speech Recognition with Uncertain Observations . . . 40
3.6.1 Uncertainty Decoding . . . 40
3.6.2 Modified Imputation . . . 41

4 MMSE Speech Enhancement with GMM Priors 43
4.1 The Need for Supergaussian Models . . . 43
4.1.1 Prior Distributions Conditioned on the a Priori SNR . . . 44
4.1.2 Prior Distributions Conditioned on Knowledge of the Variance . . . 45
4.2 The Complex Gaussian Mixture Model . . . 46
4.2.1 Definition . . . 46
4.2.2 Properties . . . 47
4.3 Complex Gaussian Mixture Model Based Estimators . . . 48
4.3.1 Differences with Respect to the Gaussian Model . . . 48
4.3.2 Rice Mixture Posterior Distribution . . . 49
4.3.3 Rice Posterior Distribution . . . 51
4.3.4 MMSE Amplitude Estimators . . . 52
4.3.5 Estimating the Statistics of Each Mixture . . . 53
4.3.6 Relation to the Rayleigh Mixture Model Estimator . . . 54
4.4 Noise Prior conditioned by Noise Variance Estimation . . . 54

5 Fourier Domain Uncertainty Models 57
5.1 A Generalized Model of Uncertainty in the Fourier Domain . . . 58
5.1.1 The Complex Gaussian Model of Uncertainty . . . 58
5.1.2 The Non-Central Chi and Non-Central Chi-Square Models of Uncertainty . . . 59
5.2 Uncertainty Estimation based on Empirical Methods . . . 60
5.2.1 Estimation of Uncertainty for the ETSI Advanced Front-End . . . 61
5.3 Posterior Distribution as a Measure of Uncertainty . . . 63
5.3.1 MMSE-STSA Posterior Distribution . . . 63
5.3.2 Wiener Filter Posterior Distribution . . . 64
5.3.3 An Approximated Posterior When Combining MMSE Amplitude Estimators with the Optimal Phase Estimator . . . 66
5.3.4 Supergaussian Measures of Uncertainty . . . 68

6 Propagation of Uncertainty through the Feature Extraction Process 69
6.1 Piecewise Uncertainty Propagation . . . 69
6.2 Uncertainty Propagation for the Mel-cepstral Features . . . 70
6.2.1 Propagation through the STSA . . . 70
6.2.2 Propagation through the Mel-Filterbank . . . 72
6.2.3 Propagation through the Logarithm of the Mel-STSA . . . 73
6.2.4 Propagation through the Discrete Cosine Transform . . . 76
6.3 Uncertainty Propagation for the RASTA-PLP Cepstral Features . . . 76
6.3.1 Propagation through the PSD . . . 76

6.3.2 Propagation through the Bark Filterbank . . . 77
6.3.3 Propagation through the Logarithm of the Bark-PSD . . . 78
6.3.4 Propagation through RASTA Filter . . . 78
6.3.5 Propagation through Pre-Emphasis, Power Law of Hearing and Exponential . . . 80
6.3.6 Propagation through the All Pole Model and LPCC Computation 81
6.4 Uncertainty Propagation for the ETSI ES 202 050 Standard . . . 81
6.4.1 Propagation of the Complex Gaussian Model Into the Time Domain 83
6.4.2 Propagation through Log-Energy Transformation . . . 85
6.4.3 Propagation through Blind Equalization . . . 86
6.4.4 Transmission and Server Side Computations . . . 86
6.5 Taking Full Covariances Into Consideration . . . 86
6.5.1 Covariances Between Features of the Same Frame . . . 86
6.5.2 Covariance Between Features of Different Frames . . . 87
6.6 Summary Table of Transformations and Assumed Distributions . . . 87
6.7 Derivations . . . 89
6.7.1 Propagation through the Exponential . . . 89

7 Experimental Results 90
7.1 Montecarlo Simulation of Uncertainty Propagation . . . 90
7.1.1 Assessing the Assumed PDFs through Histograms . . . 92
7.1.2 Accuracy of the Developed Uncertainty Propagation Algorithms . 100
7.2 Tests with the AURORA5 Database and the ETSI ES 202 050 Standard for Robust Speech Recognition . . . 104
7.2.1 AURORA5 Noisy Speech Database . . . 104
7.2.2 Implementation of Uncertainty Propagation for the ETSI Advanced Front-End . . . 104
7.2.3 Speech Recognition Results . . . 105
7.3 Tests with the AURORA5 Database and AFE-RASTA Features . . . 106
7.3.1 Speech Recognition Results . . . 107
7.4 Tests using GMM Priors for Speech and Noise . . . 108
7.4.1 Speech with Artificially Added Noise . . . 108
7.4.2 Implemented Speech Enhancement Baseline . . . 108
7.4.3 Implemented Supergaussian Estimators . . . 108
7.4.4 Speech Recognition Results . . . 111
7.5 Tests using MMSE Posterior Propagation . . . 111
7.5.1 Implemented Speech Enhancement Baseline . . . 111
7.5.2 Implemented Posterior Propagation Methods . . . 112
7.5.3 Speech Recognition Results . . . 112

8 Discussion and Conclusions 113
8.1 Discussion . . . 113
8.1.1 Piecewise Uncertainty Propagation . . . 113
8.1.2 Limitations of Modified Imputation and Uncertainty Decoding . . 114

8.1.3 Sensibility of Modified Imputation and Uncertainty Decoding to the Uncertainty Variance . . . 115
8.1.4 Improving the Advanced Front-End with Uncertainty Propagation 115
8.1.5 Posterior Propagation as MMSE-Feature Estimation . . . 116
8.1.6 Noise Prior conditioned by Noise Variance Estimation . . . 117
8.1.7 Limitations of Supergaussian Priors . . . 118
8.1.8 Supergaussian Uncertainty Propagation . . . 118
8.2 Conclusions . . . 119
8.3 Future Work . . . 120

List of Abbreviations **121**

List of Symbols **122**

Bibliography **127**

1 Overview

An automatic speech recognition (ASR) system is a device able to transcribe human speech into written words or its constituents. Such systems have a wide range of applications, from human-machine interfaces to audio indexing devices. Unfortunately, the performance of ASR systems decreases sharply when the incoming speech signal is corrupted by noises not present in the training data. Multiple approaches have been developed to try to solve this problem. Methods like parallel model combination (PMC) [Gales 1995] or vector Taylor series (VTS) model compensation [Acero 2000] attempt to dynamically compensate the model's statistical parameters to correctly recognize the noisy data. These methods, however, have high computational costs.
In recent years, uncertainty decoding, or more generally observation uncertainty techniques [Deng 2002, Droppo 2002, Kristjansson 2002], have emerged as an alternative to the previously mentioned model compensation techniques. The main idea behind this approach is to consider the speech features no longer deterministic but probabilistic or uncertain, where the uncertainty represents the loss of information that causes the mismatch between the observed speech and the trained ASR model. This uncertainty can then be combined with the statistical models of the ASR system to yield superior speech recognition performance.
This thesis explores a particular source of observation uncertainty, the one obtained from the pre-processing of the noisy signal with statistical speech enhancement techniques in the short-time Fourier transform (STFT) domain. Such speech enhancement methods are employed to improve the quality of corrupted speech signals before they are passed down either to a human listener or an ASR system. Their integration with ASR systems is however very limited until now. Typically, a corrupted version of the original clean signal is received at the microphone. A conventional speech enhancement system (Figure 1.1, left) would provide a point estimate of the spectrum of the clean speech given the corrupted speech. This estimation would then be directly passed to the ASR system by performing feature extraction of this point estimate.
As it will be demonstrated throughout this thesis, it is possible to achieve a better integration of the speech enhancement and ASR systems by replacing the point estimate by an 'uncertain' or probabilistic description of the clean signal spectrum (Figure 1.1, right). By propagating this uncertain Fourier spectrum through the feature extraction process, an uncertain description of the speech features is obtained. These uncertain features can be combined with the ASR system employing the abovementioned observation uncertainty techniques to achieve a more robust speech recognition.
This work is structured as follows: Chapter 2 introduces statistical speech enhancement in the STFT domain. Since most of the mathematical derivations in Chapter 2 play an important role in Chapters 4, 5 and 6, Section 2.3 'Derivations' has been included in this chapter with detailed descriptions of the algorithms. Chapter 3 is the introductory

chapter for conventional automatic speech recognition and automatic speech recognition under observation uncertainty. Chapter 4 presents an extension of the conventional framework for speech enhancement in the STFT domain using supergaussian a priori distributions. Chapter 5 deals with techniques for obtaining a probabilistic description of the clean signal after the speech enhancement pre-processing. Chapter 6 presents techniques for the propagation of this probabilistic information through the feature extraction process. Chapter 7 details the experimental setup employed for assessing the efficiency of the proposed algorithms and the obtained results. Chapter 8 finally contains the discussion and conclusion of the thesis.

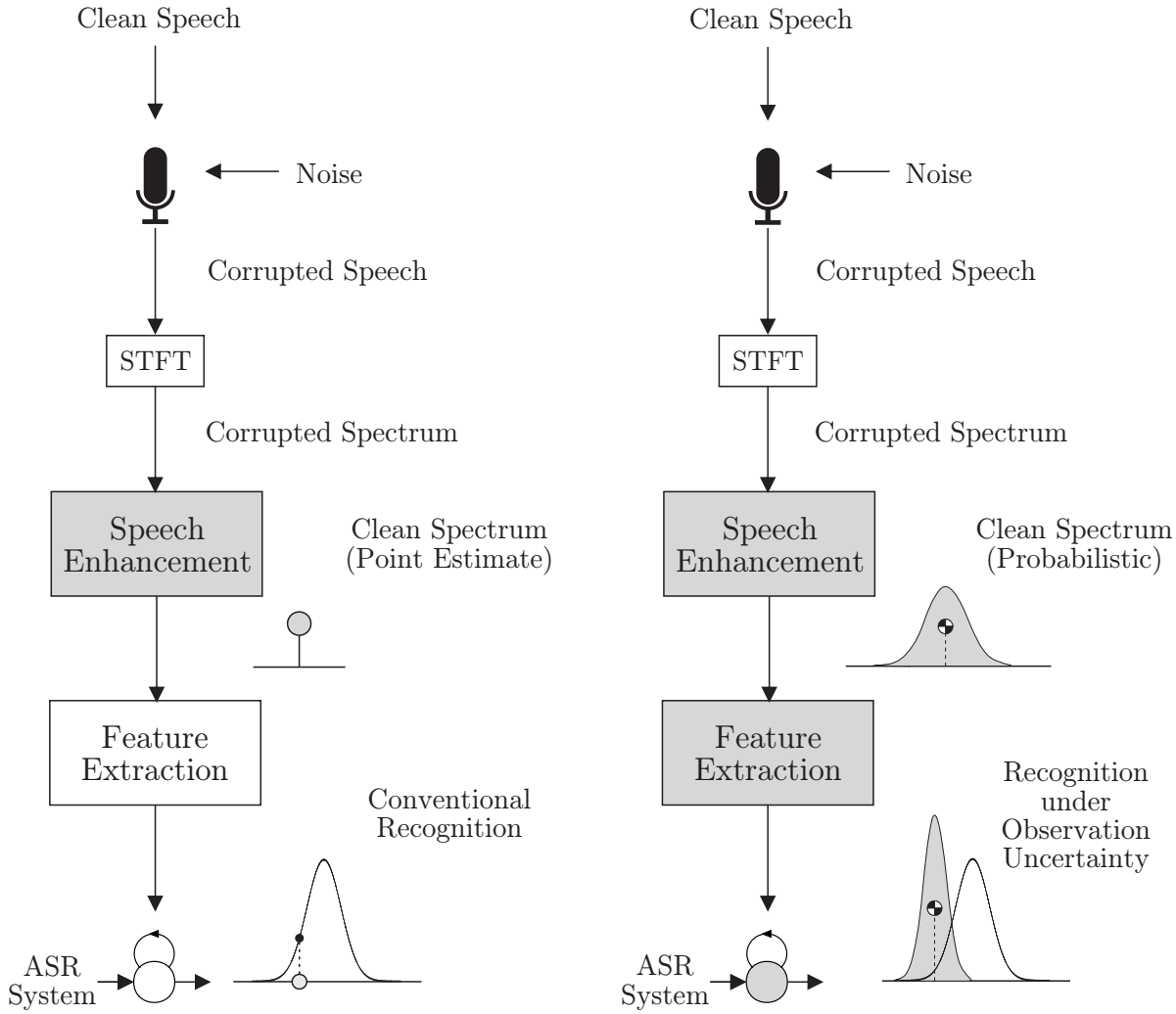

Figure 1.1: **Left:** Robust automatic speech recognition with speech enhancement preprocessing in the short-time Fourier transform (STFT) domain. **Right:** Robust automatic speech recognition with uncertainty propagation from the STFT domain into the recognition domain and use of observation uncertainty techniques.

2 Speech Enhancement for Additive Noise and Gaussian a Priori Models

Most of the classical statistical speech enhancement methods in the short-time Fourier transform (STFT) domain, like power subtraction, the Wiener filter for speech enhancement [McAulay 1980] and the Ephraim-Malah filters [Ephraim 1984, Ephraim 1985] are derived under the so called Gaussian model. This model is based on the following assumptions

1. Speech and noise are instationary, non-ergodic random processes.
2. Speech is corrupted by additive noise.
3. Speech and noise are modeled as zero mean Gaussian processes.

Figure 2.1 depicts the typical block structure of a speech enhancement system under such assumptions. The system receives a noisy signal digitized into a vector of time domain samples

$$y(t) = x(t) + d(t), \tag{2.1}$$

which contains a mixture of clean speech $x(t)$ and noise $d(t)$ and outputs an estimation of the clean speech signal $\hat{x}(t)$.

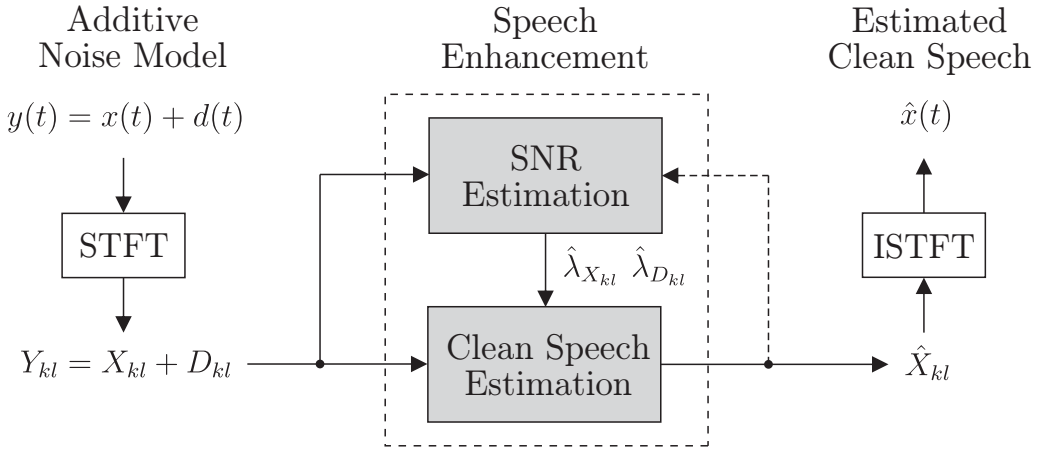

Figure 2.1: Block diagram of a speech enhancement system for speech corrupted by additive noise. The clean signal $x(t)$ is estimated from the noisy signal $y(t)$ using its STFT Y_{kl} and the estimated speech and noise variances $\hat{\lambda}_X$, $\hat{\lambda}_D$.

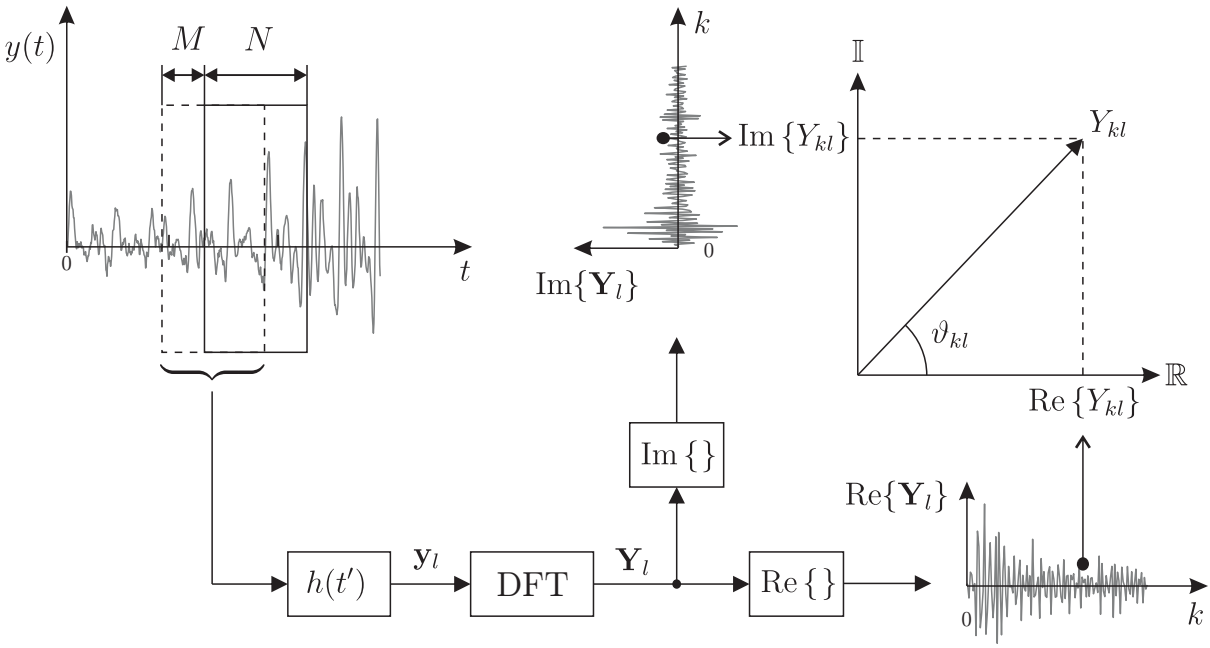

Figure 2.2: Diagram of the STFT of the noisy signal $y(t)$. A complex valued frequency representation Y_{kl} of the l^{th} frame of time domain samples $y_{t'l}$ is obtained by means of the discrete Fourier transform. The k^{th} complex valued Fourier coefficient of the l^{th} frame, Y_{kl}, is displayed on the right.

In most of the cases considered in this thesis, the speech enhancement system operates solely in the STFT domain[1], where modeling and estimation of noise and speech parameters is easier.
As displayed in Figure 2.2, the STFT consists of two steps. First the time domain signal $y(t)$ is subdivided into a set of L overlapping frames of size N. Each element of each frame is obtained as

$$y_{t'l} = y\left(t' + (l-1)M\right) h(t'), \tag{2.2}$$

where $t' \in [1 \cdots N]$ is the sample index within the frame and $l \in [1 \cdots L]$ is the frame index. Each frame is shifted by M samples with respect to the previous and weighted by a window function $h(t')$. In a second step, each frame $\mathbf{y}_l$ is transformed into the frequency domain using the discrete Fourier transform (DFT) to obtain a time-frequency representation of the signal as

$$Y_{kl} = \sum_{t'=1}^{N} y_{t'l} \exp\left(-2\pi i \frac{(t'-1)(k-1)}{N}\right), \tag{2.3}$$

where i is the imaginary unit, the index $k \in [1 \cdots K]$ corresponds to the index of each Fourier bin up to half the sampling frequency.

[1]See [Deller 1987, pp. 251-261] for a detailed description of the STFT and related methods.

Speech enhancement methods obtain an estimation of the clean signal by applying a time-frequency dependent gain function G_{kl} to obtain each corresponding estimation $\hat{X}_{kl}$ of the clean Fourier coefficient X_{kl} as

$$\hat{X}_{kl} = G_{kl} \cdot Y_{kl}. \tag{2.4}$$

Once an estimation $\hat{X}_{kl}$ of each Fourier coefficient has been obtained, the time domain signal can be recovered by using the inverse short-time Fourier transform (ISTFT), which implies inverting the DFT and using the overlap and add method (see for example [Cohen 2001, Eq. 3]).
The gain G_{kl} is obtained by combining statistical inference methods like maximum likelihood (ML), minimum mean square error (MMSE) or maximum a posteriori (MAP) estimators. The process can be considered to be divided into two main steps as depicted in Figure 2.1.

1. Clean speech estimation (Figure 2.1, lower block). An estimator gain G_{kl} is computed as a function of Y_{kl} and the variances of the noise and speech processes $\lambda_{D_{kl}}$ and $\lambda_{X_{kl}}$. Due to the assumption of instationarity and non-ergodicity, these variances change for each time and frequency element D_{kl}, X_{kl}. The ratios $|Y_{kl}|^2/\lambda_{D_{kl}}$ and $\lambda_{X_{kl}}/\lambda_{D_{kl}}$ are often used instead of the variances since they have a better numerical behavior. They are referred as signal to noise ratios (SNRs).

2. Estimation of the variances $\lambda_{X_{kl}}$ and $\lambda_{D_{kl}}$ or SNR estimation (Figure 2.1, upper block). Since the variances of speech and noise are not available, they must be estimated in a previous or parallel process. This process often relaxes the assumption of instationarity to compute the variances (i.e. by averaging over noise samples of previous frames).

2.1 Clean Speech Estimation

In order to apply any type of statistical inference technique, it is first necessary to establish models to describe the behavior of each speech and noise Fourier coefficient and their interaction. Due to the assumption of instationarity and non-ergodicity[2], each of the Fourier coefficients Y_{kl} of the noisy signal has to be modeled as a single, statistically independent, variable which is formed by the interaction of the corresponding clean speech and noise Fourier coefficients X_{kl} and D_{kl}. In the clean speech estimation block, each tuple of elements $\{Y_{kl}, X_{kl}, D_{kl}\}$ is therefore processed independently of the others, ignoring possible correlations. Consequently, and for the sake of simplicity, frequency bin and frame indices will only appear when strictly necessary in the rest of the section.

[2]Here the ergodicity assumption would imply that different time-frequency elements of the STFT matrix X_{kl}, with $k \in [1 \cdots K]$, $l \in [1 \cdots L]$, would correspond to samples of the same process, i.e a process with same statistical parameter λ_X.

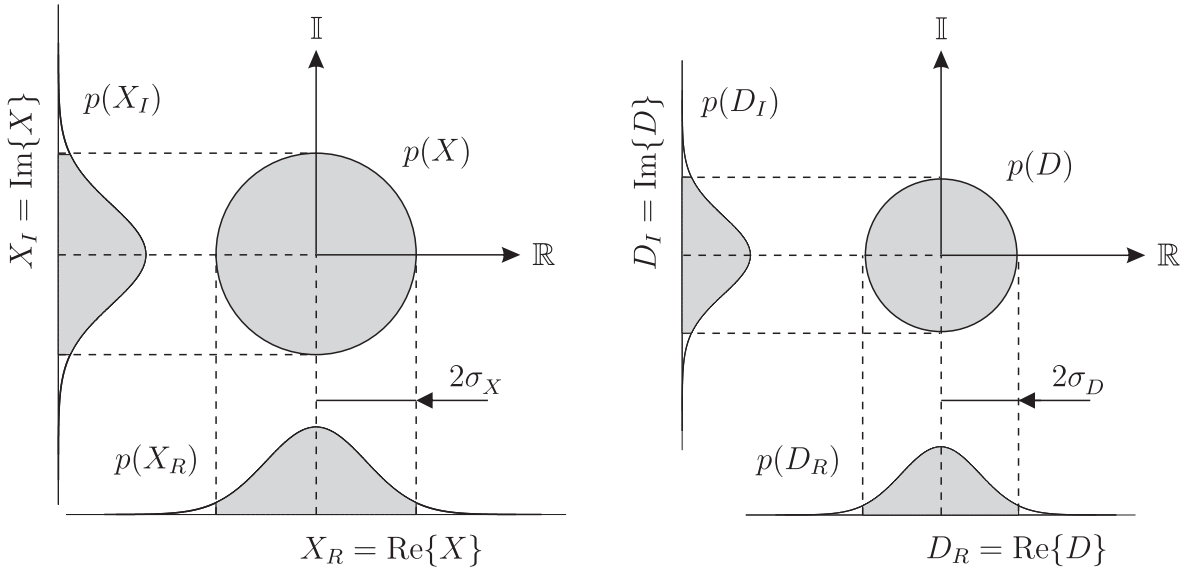

Figure 2.3: A priori distributions for a speech Fourier coefficient $p(X)$ (left) and a noise Fourier coefficient $p(D)$ (right). The Gaussian distributed real and imaginary components of X and D are displayed. The two standard deviation confidence intervals of the real and imaginary components have been shadowed. The values of the complex variable with an amplitude smaller than two component standard deviations have also been shadowed.

2.1.1 Speech and Noise a Priori Models

The first statistical models to be established are the so called a priori distributions, which summarize the previous information we have about the parameters that we wish to estimate. Under the assumed Gaussian model, the real and imaginary components of each Fourier coefficient of speech X and noise D are modeled as statistically independent Gaussian processes with mean equal to zero and variances σ_X^2 and σ_D^2 respectively [Ephraim 1984].
If we consider, for example, the real component of the clean speech and noise Fourier coefficients

$$X_R = \mathrm{Re}\{X\}, \tag{2.5}$$

$$D_R = \mathrm{Re}\{D\}, \tag{2.6}$$

their probability density functions (PDFs) would be given by

$$p(X_R) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{X_R^2}{2\sigma_X^2}\right), \tag{2.7}$$

$$p(D_R) = \frac{1}{\sqrt{2\pi\sigma_D^2}} \exp\left(-\frac{D_R^2}{2\sigma_D^2}\right), \tag{2.8}$$

where, as displayed in Figure 2.3, the corresponding imaginary components

$$X_I = \mathrm{Im}\{X\}, \tag{2.9}$$

$$D_I = \mathrm{Im}\{D\}, \tag{2.10}$$

would follow the same distribution. Taking into account this model and, by applying the definition of the complex Gaussian distribution (see Section 2.3.1 for details), we have that X and D follow the zero mean circularly symmetric complex Gaussian distributions

$$p(X) = p(X_R)p(X_I) = \frac{1}{\pi\lambda_X}\exp\left(-\frac{|X|^2}{\lambda_X}\right), \tag{2.11}$$

$$p(D) = p(D_R)p(D_I) = \frac{1}{\pi\lambda_D}\exp\left(-\frac{|D|^2}{\lambda_D}\right), \tag{2.12}$$

where $X, D \in \mathbb{C}$. The relation between the variances of the complex Fourier coefficients and the variance of their real and imaginary components is given by[3]

$$\lambda_X = E\{|X|^2\} = E\{X_R^2\} + E\{X_I^2\} = 2\sigma_X^2, \tag{2.13}$$

$$\lambda_D = E\{|D|^2\} = E\{D_R^2\} + E\{D_I^2\} = 2\sigma_D^2. \tag{2.14}$$

The assumed Gaussian model is not free from controversy. Each Fourier coefficient, as computed by Equation 2.3, can be seen as a weighted sum of random variables. If these variables were to form an strongly mixing process, the central limit theorem would justify the Gaussian assumption. This is however only asymptotically true for very large STFT frames for which sufficiently separated elements can be considered weakly dependent [Pearlman 1978, Th. 1]. Furthermore, due to the assumption of non-ergodicity it is not possible to infer the distribution describing each noise and speech Fourier coefficient from the long-term histographic analysis of signals [Ephraim 1984]. Many authors [Porter 1984, Martin 2005, Erkelens 2007] consider supergaussian distributions more adequate for the representation of speech. The ubiquity of the Gaussian model is due to the simplicity of the resulting closed-form estimators [Ephraim 2004] and the principle that, in the absence of accurate higher order information, the Gaussian model remains the optimal model to be used [Jaynes 1970]. Supergaussian models for speech and noise will be further discussed in Chapter 4.

2.1.2 Likelihood Function

The basic element of any statistical speech enhancement algorithm is the model that links the observable data with the hidden data that we wish to estimate. This is known as the likelihood function and corresponds to a conditional probability distribution function that describes the behavior of the observable data, in this case the noisy speech Fourier

[3]Note the difference between the variance of a zero mean real valued random variable $\sigma_X^2 = E\{X_R^2\}$ and a zero mean complex valued random variable $\lambda_X = E\{|X|^2\}$ [Kay 1993, pp. 500].

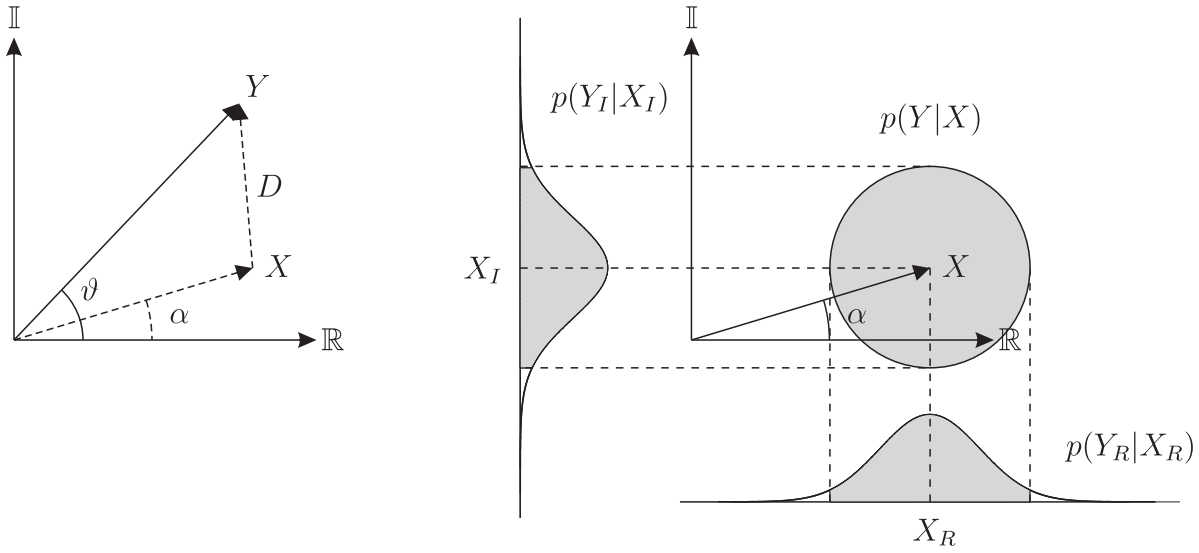

Figure 2.4: **Left:** Additive model for the noisy Fourier coefficient. The observable noisy Fourier coefficient Y, with phase ϑ, is obtained as a sum of the unknown clean signal Fourier coefficient X, with phase α, and the noise Fourier coefficient D. **Right:** Complex Gaussian likelihood function $p(Y|X)$ and Gaussian marginal distributions $p(Y_R|X_R)$, $p(Y_I|X_I)$ in cartesian coordinates. The likelihood function provides a probabilistic description of the observable data Y given the hidden data X.

coefficient Y, provided that the hidden data, that is the corresponding clean speech Fourier coefficient X, is known (see Figure 2.4). Under the assumption of additive noise given by Equation 2.1, and due to the linearity of the STFT, we have

$$Y = X + D. \tag{2.15}$$

Assuming X as known (deterministic) and D to be a random process ruled by Equation 2.12, the likelihood function will be equivalent to the probability density function of D but with mean equal to X, thus corresponding to the complex Gaussian distribution

$$p(Y|X) = \frac{1}{\pi\lambda_D} \exp\left(-\frac{|Y-X|^2}{\lambda_D}\right), \tag{2.16}$$

of mean equal to X and variance λ_D equal to that of the noise process. Note that no assumption about the prior distribution of speech was needed to obtain the likelihood function.

2.1.3 Noisy Signal Distribution

From the given model, the probability distribution $p(Y)$ of the noisy STFT coefficient can be easily computed by marginalizing (integrating out) X as

$$p(Y) = \int p(Y, X)dX = \int p(Y|X)p(X)dX. \tag{2.17}$$

However, in the case of the complex Gaussian model it is not necessary to solve this integral. Y is the result of adding the two statistically independent complex random variables defined by Equations 2.11 and 2.12. The assumption of Gaussianity of the real and imaginary component assures us therefore that the resulting distribution will remain complex Gaussian and with a variance equal to the sum of of the variances of speech and noise, hence leading to

$$p(Y) = \frac{1}{\pi(\lambda_X + \lambda_D)} \exp\left(-\frac{|Y|^2}{\lambda_X + \lambda_D}\right). \tag{2.18}$$

This distribution completes the description of the statistical model needed to perform estimation of clean speech. The next sections will be centered on MMSE estimation methods of the clean Fourier coefficient using this model.

2.1.4 Wiener Filter Estimation of Amplitude and Phase

The term Wiener filter describes a vast collection of estimators that fall into the category of linear minimum mean square error estimators (LMMSE) [Kay 1993, Chap. 12]. In general, the MMSE criterion of estimation would be to minimize the expected Euclidean error between the estimated Fourier coefficient of the clean signal $\hat{X}$ and the true Fourier coefficient of the clean signal X. This can be expressed as

$$\hat{X} = \arg\min_{\hat{X}} \left\{ E\left\{ \left\| X - \hat{X} \right\|^2 \right\} \right\}, \tag{2.19}$$

where $\|\|$ is the euclidean distance and the expectation is taken with respect to the joint distribution of X and Y. The LMMSE estimators correspond to a subset of solutions to Equation 2.19 where the estimator is constrained to be linear. Such estimators are only optimal MMSE estimators when the prior and likelihood functions are Gaussian distributed [Kay 1993, pp. 325]. Fortunately, for the complex Gaussian model presented, given by Equation 2.11 and Equation 2.16, real and imaginary components of the distribution are statistically independent and Gaussian distributed. This allows to express the expected error as the sum of the two positive terms

$$E\left\{ \left\| X - \hat{X} \right\|^2 \right\} = E\left\{ (X_R - \hat{X}_R)^2 \right\} + E\left\{ (X_I - \hat{X}_I)^2 \right\}, \tag{2.20}$$

were each term can be minimized using a LMMSE solution. Consequently an estimation of the clean Fourier coefficient X can be obtained by applying the Wiener filter to the real component

$$Y_R = \mathrm{Re}\{Y\}, \tag{2.21}$$

and the imaginary component

$$Y_I = \mathrm{Im}\{Y\}, \tag{2.22}$$

of the noisy Fourier coefficient Y separately and solving the minimization problems

$$\hat{X}_R^{\mathrm{MMSE}} = \arg\min_{\hat{X}_R} \left\{ E\left\{ (X_R - \hat{X}_R)^2 \right\} \right\}, \tag{2.23}$$

and

$$\hat{X}_I^{\mathrm{MMSE}} = \arg\min_{\hat{X}_I} \left\{ E\left\{ (X_I - \hat{X}_I)^2 \right\} \right\}. \tag{2.24}$$

This leads to the following estimator of the clean Fourier coefficient

$$\hat{X}^{\mathrm{W}} = \hat{X}_R^{\mathrm{MMSE}} + i\hat{X}_I^{\mathrm{MMSE}}, \tag{2.25}$$

The minimization problems in Equations 2.23 and 2.24 can be directly solved by applying the LMMSE solution in [Kay 1993, Eq. 12.56]. However, for the purpose of extending this introduction to the MMSE estimators, the more general MMSE solution will be explained instead. Let us consider the MMSE estimation of the real component X_R. The solution to a MMSE minimization problem, like the one in Equation 2.23, corresponds to the center of probabilistic density or expectation $E\{X_R|Y_R\}$ of the posterior distribution $p(X_R|Y_R)$ (see Section 2.3.4 for the demonstration). This expectation can be obtained from the prior $p(X_R)$ and the likelihood $p(Y_R|X_R)$ distributions by applying the Bayes theorem as

$$\hat{X}_R^{\mathrm{MMSE}} = E\{X_R|Y_R\} = \int X_R p(X_R|Y_R) dX_R = \int X_R \frac{p(Y_R|X_R)p(X_R)}{p(Y_R)} dX_R. \tag{2.26}$$

If we also solve the equivalent problem for the imaginary part X_I, the Wiener filter solution can be expressed as

$$\hat{X}^{\mathrm{W}} = \hat{X}_R^{\mathrm{MMSE}} + i\hat{X}_I^{\mathrm{MMSE}} = E\{X_R|Y_R\} + iE\{X_I|Y_I\}. \tag{2.27}$$

After solving Equation 2.26 and its imaginary counterpart (see Section 2.3.5) we obtain

$$\hat{X}^{\mathrm{W}} = \frac{\lambda_X}{\lambda_X + \lambda_D} Y_R + i\frac{\lambda_X}{\lambda_X + \lambda_D} Y_I = \frac{\lambda_X}{\lambda_X + \lambda_D} Y. \tag{2.28}$$

This is normally expressed as

$$\hat{X}^{\mathrm{W}} = G(\xi)^{\mathrm{W}} \cdot Y, \tag{2.29}$$

where

$$G(\xi)^{\mathrm{W}} = \frac{\xi}{1+\xi}, \tag{2.30}$$

is the a gain function and ξ corresponds to the a priori signal to noise ratio defined in [McAulay 1980] as

$$\xi = \frac{\lambda_X}{\lambda_D}. \tag{2.31}$$

2.1.5 Ephraim-Malah Amplitude Estimators

Ephraim-Malah filters [Ephraim 1984, Ephraim 1985] are MMSE estimators that take advantage of the fact that the amplitude of the speech plays a much more important role in the audible quality than its phase [Wang 1982]. Consequently the phase can be considered a nuisance parameter and can be integrated from the model [McAulay 1980]. Using this principle, Ephraim and Malah first derived an estimator of the amplitude of the clean Fourier coefficient $A = |X|$ termed minimum mean square error short-time spectral amplitude (MMSE-STSA) estimator [Ephraim 1984]. Let

$$X = Ae^{i\alpha}, \tag{2.32}$$

be the Fourier coefficient of the clean signal expressed in polar coordinates. The phase α of X can be marginalized (integrated out) from the speech prior $p(X)$ in Equation 2.11 and the likelihood function $p(Y|X)$ in Equation 2.16, leading to the amplitude prior (see Section 2.3.2)

$$p(A) = \int_0^{2\pi} p(A, \alpha) d\alpha = \frac{2A}{\lambda_X} \exp\left(-\frac{A^2}{\lambda_X}\right), \tag{2.33}$$

and likelihood distribution (see Section 2.3.6)

$$p(Y|A) = \int_0^{2\pi} p(Y|A, \alpha) p(\alpha) d\alpha = \frac{1}{\pi\lambda_D} \exp\left(-\frac{|Y|^2 + A^2}{\lambda_D}\right) I_0\left(\frac{2|Y|A}{\lambda_D}\right), \tag{2.34}$$

where I_0 is the modified Bessel function of order zero [Gradshteyn 2007, pp. 910-949]. Once the phase has been marginalized, the MMSE estimator of the short-time spectral amplitude (MMSE-STSA) can be obtained by solving the standard MMSE minimization problem as

$$\hat{A}^{\mathrm{MMSE}} = \arg\min_{\hat{A}} \left\{ E\left\{ (A - \hat{A})^2 \right\} \right\}. \tag{2.35}$$

Equivalently to the case of the Wiener filter, the solution to this minimization problem corresponds to the center of probabilistic density of the posterior distribution $p(A|Y)$, that can be obtained via the Bayes theorem as

$$\hat{A}^{\mathrm{MMSE}} = E\{A|Y\} = \int A \frac{p(Y|A)p(A)}{p(Y)} dA. \tag{2.36}$$

The solution of the above integral, described in Section 2.3.6, leads to the MMSE-STSA estimator

$$\hat{A}^{\mathrm{MMSE}} = \Gamma(1.5) \frac{\sqrt{\nu}}{\gamma} \exp\left(-\frac{\nu}{2}\right) \left[(1+\nu) I_0\left(\frac{\nu}{2}\right) + \nu I_1\left(\frac{\nu}{2}\right) \right] |Y|, \tag{2.37}$$

where Γ is the Gamma function [Gradshteyn 2007, pp. 892-902], I_0 and I_1 are the modified Bessel function of order zero and one respectively and the parameter ν corresponds to

$$\nu = \frac{\xi}{1+\xi}\gamma. \tag{2.38}$$

The a priori signal to noise ratio ξ was defined in Equation 2.31 and

$$\gamma = \frac{|Y|^2}{\lambda_D}, \tag{2.39}$$

is termed a posteriori signal to noise ratio after [McAulay 1980]. This estimator is completed using the optimal phase estimator of the clean Fourier coefficient X, which is demonstrated to correspond to the phase of the noisy signal [Ephraim 1984] given by

$$\exp\left(i\hat{\alpha}\right) = \frac{Y}{|Y|}. \tag{2.40}$$

Note that this phase estimator is also coincident with the phase estimator of the Wiener filter in Equation 2.28 since the estimator has the same phase as Y. By replacing the phase estimator in Equation 2.40 and the amplitude estimator in Equation 2.37 into Equation 2.32, the estimator gain function realizes

$$G(\xi,\gamma)^{\text{MMSE-STSA}} = \Gamma\left(1.5\right)\frac{\sqrt{\nu}}{\gamma}\exp\left(-\frac{\nu}{2}\right)\left[(1+\nu)I_0\left(\frac{\nu}{2}\right)+\nu I_1\left(\frac{\nu}{2}\right)\right]. \tag{2.41}$$

An improved version of this algorithm was obtained in [Ephraim 1985] by estimating the logarithm of the short-time spectral amplitude (LSA) instead of the amplitude, which reflects better how humans process speech non-linearly. The resulting estimator is termed minimum mean square error log-spectral amplitude estimator (MMSE-LSA) and corresponds to the solution of

$$\widehat{\log A}^{\text{MMSE}} = \arg\min_{\widehat{\log A}}\left\{E\left\{\left(\log A - \widehat{\log A}\right)^2\right\}\right\}, \tag{2.42}$$

which equivalently to Equation 2.35 has the solution

$$\widehat{\log A} = E\{\log A|Y\}, \tag{2.43}$$

and leads hence to

$$\hat{A}^{\text{MMSE-LSA}} = \exp\left(E\{\log A|Y\}\right). \tag{2.44}$$

Despite the fact that the posterior $p(\log A|Y)$ is unknown, a closed form solution for its expectation $E\{\log A|Y\}$ can be found in terms of the moment of μ^{th} order $E\{A^\mu|Y\}$ of the posterior distribution of the amplitude $p(A|Y)$ by using the properties of the moment generating function. Let X be a random variable and $M_X(\mu)$ its moment generating function, then it holds

$$E\{X^m\} = \frac{d^m}{d\mu^m}M_X(\mu)\Big|_{\mu=0} = \frac{d^m}{d\mu^m}E\left\{e^{\mu X}\right\}\Big|_{\mu=0}. \tag{2.45}$$

Consequently, considering $X = \log A$ and $m = 1$, Equation 2.44 can then be expressed as

$$\hat{A}^{\text{MMSE-LSA}} = \exp\left(\frac{d}{d\mu} E\{A^{\mu}|Y\}\Big|_{\mu=0}\right), \tag{2.46}$$

where the μ^{th} moment can be obtained similarly to the first moment in Equation 2.44 (see Section 2.3.7 for a complete derivation). Using the phase estimator in Equation 2.40, the gain of the new estimator becomes

$$G(\xi,\gamma)^{\text{MMSE-LSA}} = \frac{\xi}{1+\xi}\exp\left(\frac{1}{2}\int_{\nu}^{\infty}\frac{e^{-t}}{t}dt\right), \tag{2.47}$$

where the term in the exponential corresponds to the exponential integral function [Gradshteyn 2007, pp. 883-891].

2.1.6 Incorporation of Speech Presence Uncertainty in Speech Enhancement

All of the derived algorithms were obtained under the assumption of speech presence. In a noisy signal there are however periods in which speech is not present and therefore the computed gain function might be suboptimal for such periods. A possible solution for this is to derive the estimator considering two different states, speech presence H_1 and speech absence H_0 [McAulay 1980] with their corresponding probabilities. For the MMSE-STSA estimator Equation 2.36, the posterior expectation can then be expressed as

$$\hat{A}^{\text{MMSE}} = E\{A|Y\} = E\{A|Y,H_1\}p(H_1|Y) + E\{A|Y,H_0\}p(H_0|Y), \tag{2.48}$$

or equivalently, considering the gain function

$$G^{\text{MMSE-STSA}} = \left(p\cdot G_{H_1} + (1-p)\cdot G_{H_0}\right), \tag{2.49}$$

where $p = p(H_1|Y)$ is the speech presence probability computed from Y, G_{H_1} is the gain derived under speech presence as done in previous sections and G_{H_0} is the corresponding attenuation when only noise is present, which is normally set to low values $(0.1 - 0.01)$. The same procedure can be applied to the Wiener filter since the probability for each component is the same. The case of the MMSE-LSA estimator is slightly different due to the form of the estimator. Considering Equation 2.44 we have

$$\begin{aligned}\hat{A}^{\text{MMSE-LSA}} = \exp\left(E\{\log A|Y,H_1\}p(H_1|Y) + E\{\log A|Y,H_0\}p(H_0|Y)\right) = \\ \exp\left(E\{\log A|Y,H_1\}\right)^{p(H_1|Y)}\exp\left(E\{\log A|Y,H_0\}\right)^{p(H_0|Y)},\end{aligned} \tag{2.50}$$

which results in a geometrical average of both gains

$$G^{\text{MMSE-LSA}} = G_{H_1}^{p} G_{H_0}^{(1-p)}. \tag{2.51}$$

This is known as the optimally modified MMSE-LSA estimator gain (OM-LSA) [Cohen 2001].

2.2 SNR Estimation

The previous section dealt with the estimation of the unknown clean signal $x(t)$ from the observable noisy signal $y(t)$ by statistical inference. The formulated estimators where derived under the assumption that the noise and speech variances λ_X, λ_D were known. This information is however not available in real cases and also has to be estimated from the observable information. Instead of λ_X and λ_D, the a priori ξ and a posteriori γ signal to noise ratios, defined in Equations 2.31 and 2.39, are usually estimated due to their better numerical behavior. The conventional approach is to estimate the variance of the noise process λ_D first and derive the variance of the clean λ_X process or a priori SNR ξ from it.

2.2.1 Voice Activity Detection Based Noise Power Estimation

Throughout Section 2.1 it was assumed that speech and noise were instationary, non-ergodic processes. This implied that each Fourier coefficient of the clean speech X_{kl} was estimated from the corresponding Fourier coefficient of the noisy signal Y_{kl}, ignoring other Fourier coefficients $Y_{k'l'}$ adjacent in time and frequency. This assumption is now relaxed by assuming that the noise is a slowly varying process, which allows the estimation of the variance $\lambda_{D_{kl}}$ of D_{kl} (see Equation 2.14) by averaging over past frames containing only noise [Boll 1979]. In practice this estimation is performed by using recursive averaging as [Cohen 2001]

$$\hat{\lambda}_{D_{kl}} = \alpha\hat{\lambda}_{D_{kl-1}} + (1-\alpha)|D_{kl}|^2, \tag{2.52}$$

where $\hat{\lambda}_{D_{kl-1}}$ corresponds to the estimation of the noise variance for the previous frame. The parameter α is termed forgetting factor and is related to the assumed degree of stationarity of the random process D_{kl}. The typical values for α range from 0.85 to 0.99[4].

The use of recursive averaging requires a voice activity detection system (VAD) to determine if speech is present in a given frame of the the noisy signal. Assuming that such a system is already implemented, let H_0 and H_1 denote the speech absence and speech presence hypothesis in the l^{th} frame, respectively. The estimation rule for the noise process variance is given by

$$\hat{\lambda}_{D_{kl}} = \begin{cases} \alpha\hat{\lambda}_{D_{kl-1}} + (1-\alpha)|Y_{kl}|^2 & \text{if} \quad H_0 \text{ is true} \\ \\ \hat{\lambda}_{D_{kl-1}} & \text{if} \quad H_1 \text{ is true.} \end{cases} \tag{2.53}$$

A possible improvement to his procedure is to determine the speech presence and absence indicators H_1 and H_0 individually for each frequency or/and in a probabilistic manner. This adds more flexibility to the estimation procedure and allows to estimate noise from regions of the spectrum where speech is not present or present with a low probability. By considering the probabilities of the two hypotheses $p(H_0|Y)$ and $p(H_1|Y) = 1-p(H_0|Y)$,

[4] It is also possible to determine the forgetting factor using some basic assumptions, see for example [Martin 2001].

it is possible to join the two estimators in Equation 2.53 into one single average estimator as [Cohen 2003]

$$\hat{\lambda}_{D_{kl}} = p(H_0|Y)\left(\alpha\hat{\lambda}_{D_{kl-1}} + (1-\alpha)|Y_{kl}|^2\right) + p(H_1|Y)\hat{\lambda}_{D_{kl-1}}, \tag{2.54}$$

or equivalently

$$\hat{\lambda}_{D_{kl}} = \tilde{\alpha}\hat{\lambda}_{D_{kl-1}} + (1-\tilde{\alpha})|Y_{kl}|^2, \tag{2.55}$$

where

$$\tilde{\alpha} = \alpha + (1-\alpha)p(H_1|Y). \tag{2.56}$$

Various methods can be used to compute the speech presence probability. The method employed in this thesis is the improved minima controlled recursive averaging estimator (IMCRA) [Cohen 2003], which obtains a probability of speech presence for each time-frequency element based on the principle of minimum statistics [Martin 2001]. Since the particular derivation of the IMCRA method has not been taken into consideration for the algorithms developed in this work, it will not be further discussed.

2.2.2 Decision Directed a Priori Signal to Noise Ratio Estimation

Once the noise Fourier coefficient variance $\lambda_{D_{kl}}$ has been estimated, the variance of the clean Fourier coefficient $\lambda_{X_{kl}}$ can be easily estimated from it. The most simple way to do this is applying the power subtraction estimator [McAulay 1980] to each time frequency element as (see Section 2.3.3 for details)

$$\hat{\lambda}_{X_{kl}} = \max\{|Y_{kl}|^2 - \hat{\lambda}_{D_{kl}}, 0\}, \tag{2.57}$$

or equivalently, after dividing by the positive quantity $\hat{\lambda}_{D_{kl}}$ and applying Equation 2.31

$$\hat{\xi}_{kl} = \max\{\gamma_{kl} - 1, 0\}, \tag{2.58}$$

where ξ_{kl} and γ_{kl} are the a priori and a posteriori signal to noise ratios defined in Equations 2.31 and 2.39 respectively.
This estimator corresponds to the maximum likelihood estimator of λ_X for Equation 2.18, modified with a flooring at zero to ensure the non-negativity of the estimate. This approach, which constitutes a very simple clean speech estimator itself, has the main disadvantage of generating highly distorting residual noises. Isolated amplitude estimation errors in the spectrum result in sinusoidal tones in the time domain, known as musical noise, which seriously degrade the quality of the speech estimation. A possible improvement of this approach is the so called 'decision directed approach' developed by Ephraim and Malah [Ephraim 1984]. This method uses recursive time smoothing in a similar way as it was done with the noise estimation, making the resultant estimator less prone to musical noise. Although multiple variations of this formula exist, it is usually formulated as

$$\hat{\xi}_{kl} = \alpha \cdot \frac{G_{kl-1}^2 |Y_{kl-1}|^2}{\hat{\lambda}_{D_{kl}}} + (1-\alpha) \max\{\gamma_{kl} - 1, 0\}. \tag{2.59}$$

The first summand corresponds to an estimation of ξ_{kl} using the estimation of the Fourier coefficient of the clean signal $\hat{X}_{kl-1}$, of the previous time frame, as

$$\hat{\lambda}_{X_{kl-1}} = G_{kl-1}^2 |Y_{kl-1}|^2 = |\hat{X}_{kl-1}|^2, \tag{2.60}$$

where the gain function G_{kl-1} can be obtained from Equations 2.30, 2.41 and 2.47. This implies assuming that the signal does not change much from frame to frame.
This estimation is linearly combined with the conventional power subtraction estimator given in Equation 2.57 using the parameter α. Usually, an additional floor for the a priori signal to noise ratio estimation, $\xi_{\min}$, is set to control the maximum attenuation possible. The typical values for $\xi_{\min}$ range from $-30dB$ to $-10dB$. The parameter α is normally set to high values from 0.85 to 0.99, to avoid the musical noise generated by the maximum likelihood estimator. This however reduces the ability of the algorithm to adapt to abrupt changes in the signals and introduces an additional estimation error in the enhanced speech. This parameter is often considered to control the trade-off between residual noise and speech distortion.

2.3 Derivations

2.3.1 Complex Gaussian Distribution

Let $X \in \mathbb{C}$ be a complex number of real and imaginary components X_R and X_I respectively. If we assume these components to follow two statistically independent Gaussian distributions of means μ_R and μ_I with common standard deviation σ_X, then

$$X = X_R + iX_I, \tag{2.61}$$

where i is the imaginary unit, follows a circularly symmetric[5] complex Gaussian distribution of mean $\mu = \mu_R + i\mu_I$ and variance $2\sigma_X^2$. The probability density function (PDF) of this distribution can be easily computed by taking into account the statistical independence of real and imaginary components as follows:

$$\begin{aligned} p(X) = p(X_R) \cdot p(X_I) = \\ \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{(X_R-\mu_R)^2}{2\sigma_X^2}\right) \cdot \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{(X_I-\mu_I)^2}{2\sigma_X^2}\right) = \\ \frac{1}{2\pi\sigma_X^2} \exp\left(-\frac{(X_R-\mu_R)^2 + (X_I-\mu_I)^2}{2\sigma_X^2}\right). \end{aligned} \tag{2.62}$$

Re-arranging the function parameters as $\mu_X = \mu_R + i\mu_I$ and $\lambda_X = 2\sigma_X^2$, the expression usually found in the speech enhancement literature can be obtained:

[5]The condition of circular symmetry arises from the fact that real and imaginary components are statistically independent and have equal variance.

$$p(X) = \frac{1}{\pi\lambda_X} \exp\left(-\frac{|X-\mu_X|^2}{\lambda_X}\right), \tag{2.63}$$

where the parameters

$$\mu_X = E\{X\}, \tag{2.64}$$

and

$$\lambda_X = E\left\{|X-\mu_X|^2\right\} = E\left\{|X|^2\right\} - |\mu_X|^2, \tag{2.65}$$

correspond to the mean and variance of the complex valued random variable respectively [Kay 1993, pp. 500]. In the case of the Fourier coefficients modeled in Equations 2.11 and 2.12, the Gaussian distributions are considered to be zero mean ($\mu_X = 0$) and hence the formulation

$$p(X) = \frac{1}{\pi\lambda_X} \exp\left(-\frac{|X|^2}{\lambda_X}\right), \tag{2.66}$$

applies.

2.3.2 Rice Distribution

Let $X \in \mathbb{C}$ be the complex Gaussian variable previously defined, then its amplitude

$$A = |X|, \tag{2.67}$$

follows the Rice distribution of PDF

$$p(A) = \frac{2A}{\lambda_X} \exp\left(-\frac{A^2+|\mu_X|^2}{\lambda_X}\right) I_0\left(\frac{2A|\mu_X|}{\lambda_X}\right), \tag{2.68}$$

where I_0 is the modified Bessel function of order zero [Gradshteyn 2007, pp. 910-949]. In order to derive this PDF it is first necessary to express X in polar coordinates as

$$X = Ae^{i\alpha}, \tag{2.69}$$

where A and α correspond to the amplitude and phase of X respectively. The corresponding differential probability element yields

$$dP_X = p(X)dX = p(Ae^{i\alpha})dAAd\alpha = p(A,\alpha)dAd\alpha. \tag{2.70}$$

The phase α can be marginalized from $p(A,\alpha)$ by solving the integral

$$p(A) = \int_0^{2\pi} p(A,\alpha)d\alpha = \int_0^{2\pi} \frac{1}{\pi\lambda_X} \exp\left(-\frac{\left|Ae^{i\alpha}-|\mu_X|e^{i\theta}\right|^2}{\lambda_X}\right) Ad\alpha, \tag{2.71}$$

where the mean μ_X of the complex Gaussian distribution has also been expressed in polar coordinates as

$$\mu_X = |\mu_X| e^{i\theta}. \tag{2.72}$$

The integral in Equation 2.71 can be solved by first using the cosine theorem to re-express the numerator of the exponential as

$$p(A) = \int_0^{2\pi} \frac{A}{\pi\lambda_X} \exp\left(-\frac{A^2 + |\mu_X|^2 - 2A|\mu_X|\cos(\alpha-\theta)}{\lambda_X}\right) d\alpha. \tag{2.73}$$

Since the integration is over a full cycle, the θ inside the cosine has no influence and can therefore be ignored. Taking constant factors out of the integral we obtain

$$p(A) = \frac{A}{\pi\lambda_X} \exp\left(-\frac{A^2 + |\mu_X|^2}{\lambda_X}\right) \int_0^{2\pi} \exp\left(\frac{2A|\mu_X|\cos(\alpha)}{\lambda_X}\right) d\alpha. \tag{2.74}$$

The remaining integral corresponds to the integral representation of the 0^{th} order modified Bessel function [Gradshteyn 2007, Eq. 8.431.5]

$$I_0(z) = \frac{1}{2\pi} \int_0^{2\pi} \exp(z\cos x) dx, \tag{2.75}$$

multiplied by 2π.
By using this expression with $z = \frac{2A|\mu_X|}{\lambda_X}$ the desired expression in Equation 2.68

$$p(A) = \frac{2A}{\lambda_X} \exp\left(-\frac{A^2 + |\mu_X|^2}{\lambda_X}\right) I_0\left(\frac{2A|\mu_X|}{\lambda_X}\right), \tag{2.76}$$

is obtained.
There are two particular cases of the Rice distribution which are of interest for this thesis. Let

$$r = \frac{|\mu_X|^2}{\lambda_X}, \tag{2.77}$$

be defined as the ratio of the Rice distribution.
When r tends to zero, the Rice distribution tends to the Rayleigh distribution

$$p(A) = \frac{2A}{\lambda_X} \exp\left(-\frac{A^2}{\lambda_X}\right). \tag{2.78}$$

This is for example the case of the amplitude of the statistical model used for the Fourier coefficients of speech and noise (see Equations 2.11 and 2.12). Interestingly, for a zero mean circularly symmetric complex Gaussian distribution, the phase can also be easily integrated out of Equation 2.66, obtaining an uniform distribution of the phase

$$p(\alpha) = \int_0^{\infty} p(A,\alpha) dA = \int_0^{\infty} \frac{A}{\pi\lambda_X} \exp\left(-\frac{|Ae^{i\alpha}|^2}{\lambda_X}\right) dA = \frac{1}{2\pi} \int_0^{\infty} \frac{2A}{\lambda_X} \exp\left(-\frac{A^2}{\lambda_X}\right) dA = \frac{1}{2\pi}, \tag{2.79}$$

where the last equality is obtained by noting that the integral equals one since it corresponds to the integration of the Rayleigh PDF over its domain. Furthermore, since the product of the marginal distribution of amplitude $p(A)$ in Equation 2.78 and phase $p(\alpha)$ in Equation 2.79 equals the joint distribution in Equation 2.70

$$p(A)p(\alpha) = p(A, \alpha), \tag{2.80}$$

amplitude and phase under this model can be considered statistically independent.
Another case of interest is when the ratio r tends to infinity, a case in which the Rice distribution approximately converges to the Gaussian distribution [Gudbjartsson 1995]

$$p(A) \approx \frac{1}{\sqrt{\pi\lambda_X}} \exp\left(-\frac{\left(A - \sqrt{|\mu_X|^2 + \lambda_X/2}\right)^2}{\lambda_X} \right). \tag{2.81}$$

2.3.3 Derivation of the Power Subtraction Estimator

The distribution of the noisy Fourier coefficient Y, given in Equation 2.18 corresponds to

$$p(Y) = \frac{1}{\pi(\lambda_X + \lambda_D)} \exp\left(-\frac{|Y|^2}{\lambda_X + \lambda_D} \right). \tag{2.82}$$

This distribution can be used to obtain an estimate of the variance of the clean Fourier coefficient λ_X, given the variance of the noisy Fourier coefficient λ_D, by using the maximum likelihood method (ML) [Kay 1993, Chap. 7]. A ML estimation of λ_X can be obtained as

$$\hat{\lambda}_X^{\mathrm{ML}} = \arg\max_{\lambda_X} \{p(Y)\}. \tag{2.83}$$

Since the logarithm is a continuous strictly increasing function, the solution to Equation 2.83 can also be obtained by maximizing the log-likelihood

$$\log\left(p(Y)\right) = -\log\left(\pi(\lambda_X + \lambda_D)\right) - \frac{|Y|^2}{\lambda_X + \lambda_D}, \tag{2.84}$$

which has a simpler expression.
In order to solve Equation 2.83, the derivative of Equation 2.84 with respect to λ_X is computed and set equal to zero

$$-\frac{1}{(\hat{\lambda}_X^{\mathrm{ML}} + \lambda_D)} + \frac{|Y|^2}{(\hat{\lambda}_X^{\mathrm{ML}} + \lambda_D)^2} = 0. \tag{2.85}$$

which leads to

$$\hat{\lambda}_X^{\mathrm{ML}} = |Y|^2 - \lambda_D. \tag{2.86}$$

To prevent the variance estimate from being negative, a flooring at zero is used

$$\hat{\lambda}_X^{\mathrm{ML}} = \max\{|Y|^2 - \lambda_D, 0\}. \tag{2.87}$$

2.3.4 Derivation of the MMSE Criterion for Estimation

Without loss of generality, we will consider here the estimation $\hat{X}_R$ of the real component of the Fourier coefficient of the STFT of the clean signal X_R, from the real component of the Fourier coefficient of the STFT of the noisy signal Y_R, to exemplify the conventional MMSE minimization problem. This implies solving the minimization problem

$$\hat{X}_R^{\text{MMSE}} = \arg\min_{\hat{X}_R} \left\{ E\left\{ (X_R - \hat{X}_R)^2 \right\} \right\}, \tag{2.88}$$

which can be done by noting that the expectation to be minimized can be expressed in terms of the joint probability distribution of X_R and Y_R as

$$E\left\{ (X_R - \hat{X}_R)^2 \right\} = \int\int (X_R - \hat{X}_R)^2 p(X_R, Y_R) dX_R dY_R. \tag{2.89}$$

By decomposing the joint distribution $p(X_R, Y_R)$ into the product of the posterior distribution $p(X_R|Y_R)$ and the data prior $p(Y_R)$, the expected estimation error can be decomposed as follows

$$E\left\{ (X_R - \hat{X}_R)^2 \right\} = \int \left(\int (X_R - \hat{X}_R)^2 p(X_R|Y_R) dX_R \right) p(Y_R) dY_R, \tag{2.90}$$

which yields

$$E\left\{ (X_R - \hat{X}_R)^2 \right\} = \int \left(E\left\{ (X_R - \hat{X}_R)^2 | Y_R \right\} \right) p(Y_R) dY_R. \tag{2.91}$$

Since $p(Y_R)$ is zero or a positive number, the minimization problem in Equation 2.88 can be reduced to a minimization of the following expected estimation error [Kay 1993, pp. 312]

$$\hat{X}_R^{\text{MMSE}} = \arg\min_{\hat{X}_R} \left\{ E\left\{ (X_R - \hat{X}_R)^2 | Y_R \right\} \right\}. \tag{2.92}$$

By computing the derivative of this conditional expectation with respect to $\hat{X}_R$, taking into account the properties of the expectation operator and equating to zero we obtain the general form of the MMSE estimator, which corresponds to the expectation of the posterior distribution

$$\hat{X}_R^{\text{MMSE}} = E\{X_R|Y_R\}. \tag{2.93}$$

For a given probabilistic model, defined by a likelihood function $p(Y_R|X_R)$ and a hidden parameter prior $p(X_R)$, the MMSE estimate can be computed by using the Bayes theorem as

$$\hat{X}_R^{\text{MMSE}} = \int X_R p(X_R|Y_R) dX_R = \int X_R \frac{p(Y_R|X_R)p(X_R)}{p(Y_R)} dX_R, \tag{2.94}$$

where $p(Y_R)$ can be computed by marginalizing X_R from the joint distribution $p(X_R, Y_R)$ expressed in terms of the likelihood and prior distributions as

$$p(Y_R) = \int p(Y_R|X_R)p(X_R)dX_R. \tag{2.95}$$

2.3.5 Derivation of the Wiener Filter for Speech Enhancement

The likelihood function $p(Y|X)$ in Equation 2.16, the speech prior distribution $p(X)$ in Equation 2.11, and the the noisy Fourier coefficient distribution $p(Y)$ in Equation 2.18, are all complex Gaussian distributed. Applying the definition of complex Gaussian distribution given in Section 2.3.1, the corresponding distributions of the real components of the Fourier coefficients are Gaussian distributed and correspond to

$$p(X_R) = \frac{1}{\sqrt{\pi\lambda_X}} \exp\left(-\frac{X_R^2}{\lambda_X}\right), \tag{2.96}$$

$$p(Y_R|X_R) = \frac{1}{\sqrt{\pi\lambda_D}} \exp\left(-\frac{(Y_R - X_R)^2}{\lambda_D}\right), \tag{2.97}$$

$$p(Y_R) = \frac{1}{\sqrt{\pi(\lambda_D + \lambda_X)}} \exp\left(-\frac{Y_R^2}{\lambda_D + \lambda_X}\right). \tag{2.98}$$

The solution of the Wiener filter is obtained as

$$\hat{X}_R^{\text{MMSE}} = E\{X_R|Y_R\} = \int X_R \frac{p(Y_R|X_R)p(X_R)}{p(Y_R)} dX_R, \tag{2.99}$$

which by replacing Equations 2.96, 2.97 and 2.98 into Equation 2.99, expanding the square and taking the constants out of the integral leads to

$$\begin{aligned}\hat{X}_R^{\text{MMSE}} = &\frac{1}{\sqrt{\pi\frac{\lambda_X\lambda_D}{\lambda_X+\lambda_D}}} \int X_R \exp\left(-\frac{X_R^2}{\lambda_X} - \frac{(Y_R - X_R)^2}{\lambda_D} + \frac{Y_R^2}{\lambda_D + \lambda_X}\right) dX_R = \\ &\frac{\exp\left(\frac{Y_R^2}{\lambda_D+\lambda_X} - \frac{Y_R^2}{\lambda_D}\right)}{\sqrt{\pi\frac{\lambda_X\lambda_D}{\lambda_X+\lambda_D}}} \int X_R \exp\left(-\frac{\lambda_X + \lambda_D}{\lambda_X\lambda_D} X_R^2 + 2\frac{Y_R}{\lambda_D} X_R\right) dX_R.\end{aligned} \tag{2.100}$$

The remaining expression can be solved by applying [Gradshteyn 2007, Eq. 3.462.2.8]

$$\int_{-\infty}^{\infty} x^n \exp\left(-px^2 + 2qx\right) = \frac{1}{2^{n-1}p}\sqrt{\frac{\pi}{p}} \frac{d^{n-1}}{dq^{n-1}} \left(qe^{\frac{q^2}{p}}\right), \tag{2.101}$$

with $n = 1$, $q = Y_R/\lambda_D$ and $p = (\lambda_X + \lambda_D)/(\lambda_D\lambda_X)$. This leads to

$$\hat{X}_R^{\text{MMSE}} = \frac{\lambda_X}{\lambda_X + \lambda_D} Y_R. \tag{2.102}$$

The resolution for the imaginary part is identical, since the variances of the real and imaginary components are the same.

2.3.6 Derivation of the First Ephraim-Malah Filter for Speech Enhancement

Given the likelihood function of the additive noise model assumed in Equation 2.15

$$p(Y|X) = \frac{1}{\pi\lambda_D} \exp\left(-\frac{|Y-X|^2}{\lambda_D}\right), \tag{2.103}$$

the distribution of the noisy signal Y given the amplitude of the clean Fourier coefficient $A = |X|$ can be obtained, as in [McAulay 1980, Eqs. 13-16], by marginalization of the phase of the clean Fourier coefficient α. This is done using the fact that the amplitude A and phase α of X are statistically independent as well as the fact that the phase is uniformly distributed (see Section 2.3.2). This allows us to express the desired likelihood as [McAulay 1980]

$$p(Y|A) = \int_0^{2\pi} p(Y,\alpha|A)d\alpha = \int_0^{2\pi} p(Y|A,\alpha)p(\alpha)d\alpha = \frac{1}{2\pi}\int_0^{2\pi} p(Y|A,\alpha)d\alpha. \tag{2.104}$$

A very similar[6] integral to this one was already solved in the derivation of the Rice PDF. By changing the variable A to $|Y|$, $|\mu_X|$ to A and λ_X to λ_D in Equation 2.68 we obtain

$$p(Y|A) = \frac{1}{\pi\lambda_D} \exp\left(-\frac{|Y|^2 + A^2}{\lambda_D}\right) I_0\left(\frac{2|Y|A}{\lambda_D}\right). \tag{2.105}$$

Note that despite using the variable Y, this conditional distribution depends only on its modulus $R = |Y|$. In fact, the noisy phase ϑ can be integrated out from Equation 2.105 yielding a Rice distributed likelihood function relating clean and noisy amplitudes

$$p(R|A) = \int_0^{2\pi} p(R,\vartheta|A)Rd\vartheta = \frac{2R}{\lambda_D} \exp\left(-\frac{R^2 + A^2}{\lambda_D}\right) I_0\left(\frac{2RA}{\lambda_D}\right). \tag{2.106}$$

As any MMSE minimization problem (see Section 2.3.5), the solution to the MMSE-STSA estimator [Ephraim 1985] can be expressed as

$$\hat{A}^{\text{MMSE}} = E\{A|Y\} = \int_0^{\infty} A\frac{p(Y|A)p(A)}{p(Y)}dA. \tag{2.107}$$

By replacing Equation 2.105 and Equation 2.78 into this equation we obtain

$$\begin{aligned} E\{A|Y\} = \frac{1}{p(Y)}\int_0^{\infty} Ap(Y|A)p(A)dA = \\ \frac{1}{p(Y)}\int_0^{\infty} \frac{A}{\pi\lambda_D} \exp\left(-\frac{|Y|^2 + A^2}{\lambda_D}\right) I_0\left(\frac{2|Y|A}{\lambda_D}\right) \frac{2A}{\lambda_X} \exp\left(-\frac{A^2}{\lambda_X}\right) dA = \\ \frac{1}{p(Y)} \frac{2\exp\left(-\frac{|Y|^2}{\lambda_D}\right)}{\pi\lambda_D\lambda_X} \int_0^{\infty} A^2 \exp\left(-\frac{\lambda_D + \lambda_X}{\lambda_D\lambda_X}A^2\right) I_0\left(\frac{2|Y|A}{\lambda_D}\right) dA. \end{aligned} \tag{2.108}$$

[6] The only difference is the constant factor $2\pi A$ originating from the phase prior and the phase differential element.

The remaining integral corresponds to the first moment of the Rice distribution and can be solved as a particular case of the integral [Gradshteyn 2007, Eqs. 6.631.1, 8.406.3]

$$\int_0^\infty x^\mu e^{-\alpha x^2} I_0(\beta i x) dx = \frac{\Gamma\left(\frac{\mu+1}{2}\right)}{2\alpha^{\frac{\mu+1}{2}}} \Phi\left(\frac{\mu+1}{2}; 1; -\frac{\beta^2}{4\alpha}\right). \tag{2.109}$$

Solving this integral for $\mu = 2$, $\alpha = \frac{\lambda_D+\lambda_X}{\lambda_D\lambda_X}$ and $\beta = i\frac{2|Y|}{\lambda_D}$ and replacing it into Equation 2.108 we obtain

$$E\{A|Y\} = \frac{1}{p(Y)} \frac{\exp\left(-\frac{|Y|^2}{\lambda_D}\right)}{\pi\lambda_D\lambda_X} \left(\frac{\lambda_D\lambda_X}{\lambda_D+\lambda_X}\right)^{\frac{3}{2}} \Gamma(1.5)\, \Phi\left(1.5; 1; -\frac{\lambda_X}{\lambda_D+\lambda_X}\frac{|Y|^2}{\lambda_D}\right), \tag{2.110}$$

where $\Gamma()$ is the Gamma function and $\Phi()$ corresponds to the confluent hypergeometric function [Gradshteyn 2007, pp. 1022-1032]. This expression can be further simplified by using the property of $\Phi()$ given in [Gradshteyn 2007, Eq. 9.212.1]

$$\Phi(\alpha, \gamma, z) = e^z \Phi(\gamma - \alpha, \gamma, -z), \tag{2.111}$$

and using the definition of $p(Y)$ given in Equation 2.18. This leads to the expression

$$E\{A|Y\} = \left(\frac{\lambda_D\lambda_X}{\lambda_D+\lambda_X}\right)^{\frac{1}{2}} \Gamma(1.5)\, \Phi\left(-0.5; 1; -\frac{\lambda_X}{\lambda_D+\lambda_X}\frac{R^2}{\lambda_D}\right). \tag{2.112}$$

Expressing this in terms of a priori signal to noise ratio

$$\xi = \frac{\lambda_X}{\lambda_D}, \tag{2.113}$$

the a posteriori signal to noise ratio,

$$\gamma = \frac{|Y|^2}{\lambda_D}, \tag{2.114}$$

and the parameter

$$\nu = \frac{\xi}{1+\xi}\gamma, \tag{2.115}$$

we obtain

$$E\{A|Y\} = \Gamma(1.5)\, \frac{\sqrt{\nu}}{\gamma} \Phi(-0.5; 1; -\nu)\, |Y|. \tag{2.116}$$

Finally, this case of the confluent hypergeometric function can be expressed as a Laguerre polynomial or in terms of modified Bessel functions of the first and second kind, I_0 and I_1, using [Middelton 1960, Eq. A.1.14], which are more easily found in software packets or tabulated.

$$\hat{A}^{\text{MMSE}} = \Gamma(1.5)\frac{\sqrt{\nu}}{\gamma}\mathrm{L}_{\frac{1}{2}}(-\nu)|Y| =$$
$$\Gamma(1.5)\frac{\sqrt{\nu}}{\gamma}\exp\left(-\frac{\nu}{2}\right)\left[(1+\nu)I_0\left(\frac{\nu}{2}\right)+\nu I_1\left(\frac{\nu}{2}\right)\right]|Y|. \tag{2.117}$$

2.3.7 Derivation of the Second Ephraim-Malah Filter for Speech Enhancement

The MMSE-LSA estimator can be obtained by solving[7]

$$\hat{A}^{\text{MMSE-LSA}} = \exp\left(E\{\log A|Y\}\right) = \exp\left(\frac{d}{dn}E\{A^n|Y\}\Big|_{n=0}\right). \tag{2.118}$$

The n^{th} moment of the posterior distribution $p(A|Y)$ can be obtained as

$$E\{A^n|Y\} = \frac{1}{p(Y)}\int_0^\infty A^n p(Y|A)p(A)dA. \tag{2.119}$$

Similarly to Equation 2.108, this integral can be solved by applying Equation 2.109 and [Gradshteyn 2007, Eq. 9.212.1] for $\mu = n+1$, $\alpha = \frac{\lambda_D+\lambda_X}{\lambda_D\lambda_X}$ and $\beta = i\frac{2|Y|}{\lambda_D}$ leading to

$$E\{A^n|Y\} = \left(\frac{\lambda_D\lambda_X}{\lambda_D+\lambda_X}\right)^{\frac{n}{2}}\Gamma\left(\frac{n}{2}+1\right)\Phi\left(-\frac{n}{2};1;-\nu\right). \tag{2.120}$$

The derivation of Equation 2.118 can be done by taking into account that the derivate of each of the three factors in Equation 2.120 evaluated for $n=0$, equals one. Consequently, the derivative of the product evaluated at $n=0$ is equal to the sum of the derivatives of each term evaluated at $n=0$.

$$\frac{d}{dn}E\{A^n|Y\}\Big|_{n=0} =$$
$$\frac{d}{dn}\left(\frac{\lambda_D\lambda_X}{\lambda_D+\lambda_X}\right)^{\frac{n}{2}}\Big|_{n=0} + \frac{d}{dn}\Gamma\left(\frac{n}{2}+1\right)\Big|_{n=0} + \frac{d}{dn}\Phi\left(-\frac{n}{2};1;-\nu\right)\Big|_{n=0}. \tag{2.121}$$

These derivatives correspond to

$$\frac{d}{dn}\left(\frac{\lambda_D\lambda_X}{\lambda_D+\lambda_X}\right)^{\frac{n}{2}}\Big|_{n=0} = \frac{1}{2}\ln\left(\frac{\lambda_D\lambda_X}{\lambda_D+\lambda_X}\right), \tag{2.122}$$

$$\frac{d}{dn}\Gamma\left(\frac{n}{2}+1\right)\Big|_{n=0} = \Gamma\left(\frac{n}{2}+1\right)\frac{d}{dn}\ln\Gamma\left(\frac{n}{2}+1\right)\Big|_{n=0} = -\frac{c}{2}, \tag{2.123}$$

$$\frac{d}{dn}\Phi\left(-\frac{n}{2};1;-\nu\right)\Big|_{n=0} = -\frac{1}{2}\sum_{r=1}^{\infty}\frac{(-\nu)^r}{r!}\frac{1}{r}, \tag{2.124}$$

[7]Note the notation change with respect to Equation 2.45 (n instead of μ) to avoid confusion.

where Equation 2.123 is obtained from the derivative of $\ln\Gamma$, the logarithm of the gamma distribution evaluated at $n = 0$ was computed by using its series representation [Gradshteyn 2007, Eq. 8.342.1] and c is the Euler-Mascheroni constant. The derivative of the hypergeometric function has also been obtained by means of its series representation [Gradshteyn 2007, Eq. 9.210.1]. Replacing these three derivatives in Equation 2.121 and applying Equation 2.118 we obtain the following closed form solution for the MMSE-LSA estimator

$$\hat{A}^{\text{MMSE-LSA}} = \exp\left(E\{\log A|Y\}\right) = \exp\left(\frac{d}{dn}E\{A^n|Y\}\Big|_{n=0}\right) = \exp\left(\frac{1}{2}\ln\left(\frac{\lambda_D\lambda_X}{\lambda_D+\lambda_X}\right) - \frac{c}{2} - \frac{1}{2}\sum_{r=1}^{\infty}\frac{(-\nu)^r}{r!}\frac{1}{r}\right). \quad (2.125)$$

This expression can be reformulated by using the expression of the exponential integral function given by [Gradshteyn 2007, Eq. 8.214.1] and [Gradshteyn 2007, Eq. 8.211.1]

$$c + \sum_{r=1}^{\infty}\frac{(-\nu)^r}{r!}\frac{1}{r} = \text{Ei}(-\nu) - \ln(\nu) = -\int_{\nu}^{\infty}\frac{e^{-t}}{t}dt - \ln(\nu), \quad (2.126)$$

leading to

$$\hat{A}^{\text{MMSE-LSA}} = \frac{\xi}{1+\xi}\exp\left(\frac{1}{2}\int_{\nu}^{\infty}\frac{e^{-t}}{t}dt\right)|Y|, \quad (2.127)$$

where the final estimator has been expressed in terms of the a priori signal to noise ratio ξ defined in Equation 2.31. The following computationally efficient approximation developed by Martin [Martin 2004] can be used to calculate the exponential integral function

$$\int_{\nu}^{\infty}\frac{e^{-t}}{t}dt \approx \begin{cases} -2.31\log_{10}(\nu) & \text{for } \nu < 0.1 \\ -1.544\log_{10}(\nu) & \text{for } 0.1 \leq \nu \leq 1 \\ 10^{-0.52\nu-0.26} & \text{for } \nu > 1. \end{cases} \quad (2.128)$$

The exponential integral function is also already implemented into MATLAB in the expint.m function.

3 Automatic Speech Recognition with Hidden Markov Models

The task of an automatic speech recognition (ASR) system is to obtain the word, or sequence of words, that correspond to the spoken message contained in a waveform. The most successful approach to do so nowadays is to train a probabilistic model for a finite vocabulary of words using a large database of speech samples. As many other pattern recognition problems, this requires a transformation of the time domain signal into a sequence of feature vectors more suitable for training and recognition.
As depicted in Figure 3.1, hidden Markov model (HMM) based speech recognizers model each word or its constituents, in this case phonemes, by a set of states $S_1 \cdots S_6$ corresponding to different clusters of feature vectors adjacent in time. Following the Markov property of first order, the transition from one state to another is governed by a transition probability depending only on both states. In other words, let $q_l = S_i$ denote being at state i at a time frame l, the probability of being at state j in the next time frame $l+1$ is given by

$$a_{ij} = p(q_{l+1} = S_j | q_l = S_i). \tag{3.1}$$

The abstraction 'phoneme' corresponds to a multiplicity of realizations which changes for each speaker and situation (i.e. gender, accent, stress, animic state), furthermore a finite number of states might be suboptimal to describe the time evolution of the phoneme or other modeled abstractions. Consequently, each state S_j models the corresponding cluster of feature vectors using a multivariable probability distribution. That is, a state S_j corresponds not to a particular observation but a distribution of all observations with certain probability as given by

$$b_j(\mathbf{o}_l) = p(\mathbf{o}_l | q_l = S_j). \tag{3.2}$$

To provide the state emission probability of each state with enough flexibility, a multivariate Gaussian mixture model (GMM) is often used. This distribution corresponds to a weighted, normalized, sum of multivariate Gaussian distributions and realizes the formula

$$p(\mathbf{o}_l | q_l = S_j) = \sum_{m=1}^{M} W_{jm} \frac{1}{\sqrt{(2\pi)^n |\mathbf{\Sigma}_{jm}|}} \exp\left(-\frac{1}{2} (\mathbf{o}_l - \boldsymbol{\mu}_{jm}) \mathbf{\Sigma}_{jm}^{-1} (\mathbf{o}_l - \boldsymbol{\mu}_{jm})^T \right), \tag{3.3}$$

with the constraint

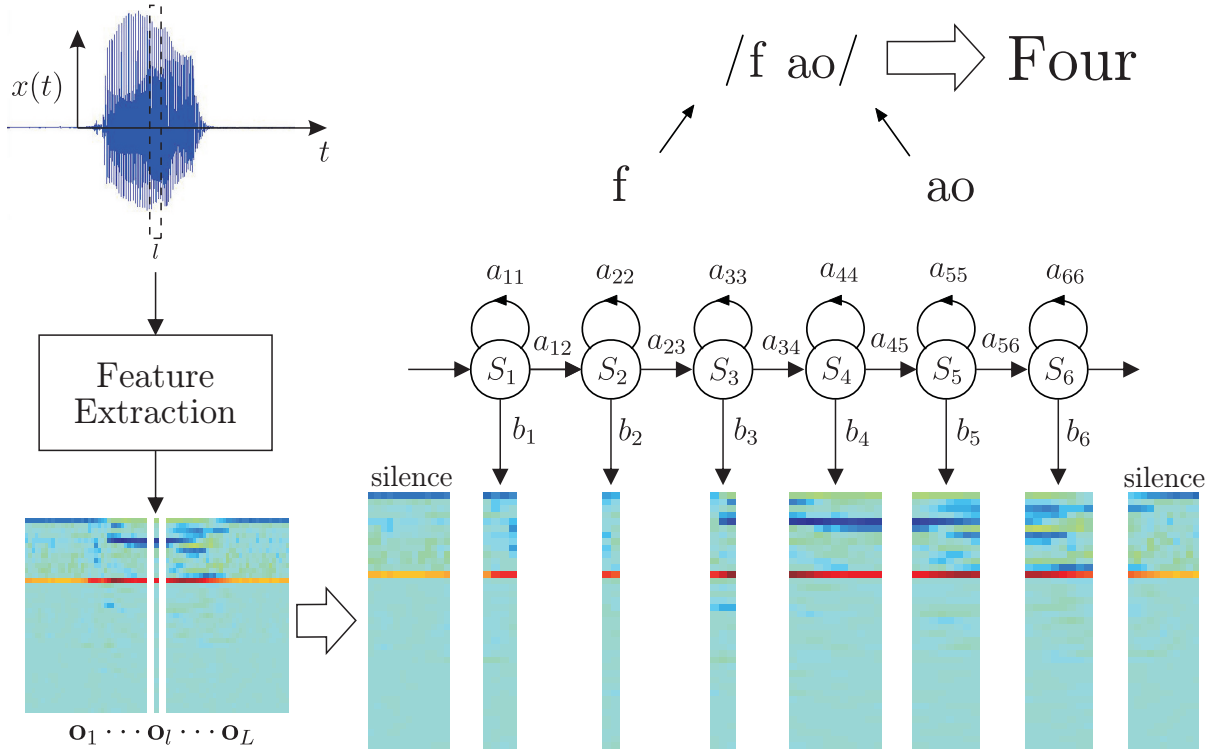

Figure 3.1: **Left:** Speech waveform $x(t)$ of the word 'Four' transformed into a matrix of features o_{il}. **Right:** Left-right hidden Markov model (HMM) based automatic speech recognition of the word 'Four' (phonetically /f ao/). The figure displays the feature vectors (frames) $\mathbf{o}_1 \cdots \mathbf{o}_L$ assigned to each HMM state (namely f $= \{S_1 \cdots S_3\}$ and ao $= \{S_4 \cdots S_6\}$) when recognition is finished. The HMM is shown as a finite state machine with state transition probabilities a_{ij} and state emission probabilities $b_j(\mathbf{o}_l)$.

$$\sum_{m=1}^{M} W_{jm} = 1, \tag{3.4}$$

and where $\boldsymbol{\mu}_{jm}$ and $\boldsymbol{\Sigma}_{jm}$ correspond to the vector mean and covariance matrix of mixture m belonging to state S_j. A HMM based ASR system then comprises the following

- A HMM for each word in the vocabulary, as well as extra models for silence and short-pause models. These are all conventional left-right HMMs similar to the one displayed in Figure 3.1.
- A language model describing the transition probabilities between words, that is, the transition probabilities between the last state of each word model and the first state of each other word model. In the absence of information, a fully connected word net can be assumed where each word is equally likely, regardless of the predecessor.

3.1 Training of an ASR System

Word and silence models can be trained individually, by using multiple samples of labeled data containing information about word or phoneme boundaries, or in embedded mode, in which composite models are trained with whole sentences. The standard algorithm for this matter is the expectation-maximization (EM) meta-algorithm [Dempster 1977]. Particularly, for a HMM structure with GMM emitting densities, the Baum-Welch EM algorithm [Rabiner 1989] is used. Word transition probabilities can be statically or dynamically computed from syntactical or semantical (i.e. context) models of language, which are not forced to be Markov based and can be trained separately from the word or phoneme models. An extensive description of the training strategies used in the experimental setups in Chapter 7 can be found in the HTK Book [Young 2006]. Since these are not relevant for this work, they will not be further discussed.

3.2 Recognition

The speech recognition task implies finding the sequence of words given a waveform, or equivalently the sequence of states $\hat{\mathbf{q}} = q_1 \cdots q_L$ given the observable feature vectors $\mathbf{o} = \mathbf{o}_1 \cdots \mathbf{o}_L$. This can be expressed, by using the Bayes theorem as a maximum a posteriori (MAP) maximization problem

$$\hat{\mathbf{q}} = \arg\max_{\mathbf{q}} \{p(\mathbf{q}|\mathbf{o})\} = \arg\max_{\mathbf{q}} \left\{ \frac{p(\mathbf{o}|\mathbf{q})p(\mathbf{q})}{p(\mathbf{o})} \right\} = \arg\max_{\mathbf{q}} \{p(\mathbf{o}, \mathbf{q})\}, \tag{3.5}$$

where the probability $p(\mathbf{o})$ is not relevant since it does not depend on $\mathbf{q}$ and the problem is reduced to finding the state sequence that maximizes the joint probability $p(\mathbf{o}, \mathbf{q})$. The solution to this problem can be obtained by employing dynamic programming techniques like the Viterbi algorithm [Rabiner 1989] or more general implementations like the token passing approach [Young 1989]. Both algorithms work under the same principle, which arises from the first order Markov characteristic of the model. Let $\delta_l(j)$ denote the maximum joint probability at time step l and state S_j, that is, the maximum probability $p(\mathbf{o}_1 \cdots \mathbf{o}_l, q_1 \cdots q_l)$ of having observed feature vectors $\mathbf{o}_1 \cdots \mathbf{o}_l$ at time step l and currently being at state S_j [Rabiner 1989, Eq. 30]

$$\delta_l(j) = \max_{q_1 \cdots q_{l-1}} \{p(\mathbf{o}_1 \cdots \mathbf{o}_l, q_1 \cdots q_l = S_j)\}. \tag{3.6}$$

Since there is only a finite number of states from which the actual state S_j could have been reached and due to the Markov property, this maximum likelihood solution can be expressed in terms of the maximum likelihood of the previous step [Rabiner 1989, Eq. 33a]. This recursion allows the computation of the maximum a posteriori probability in a simple way as

$$\delta_l(j) = \max_i \{\delta_{l-1}(i) a_{ij}\} b_j(\mathbf{o}_l), \tag{3.7}$$

where a_{ij} and b_j, with $i,\ j \in [1 \cdots J]$ correspond to the transition and emission probabilities of a trained HMM with J states as defined in Equations 3.1 and 3.2. After

processing all observations, the maximum joint probability is obtained as

$$\max_{\mathbf{q}} \{p(\mathbf{o}, \mathbf{q})\} = \max_i \{\delta_{L-1}(i) a_{ij}\} b_j(\mathbf{o}_L). \tag{3.8}$$

In absence of additional information, the algorithm is initialized at $l = 1$ by considering all initial states of all words equally probable. This can be done by creating the dummy variable

$$\delta_1(j) = \begin{cases} \frac{b_j(\mathbf{o}_1)}{N_W} & \text{if } j \text{ correspond to an initial state of a word} \\ 0 & \text{otherwise} \end{cases} \tag{3.9}$$

where N_W is the total number of words in the HMM.
To obtain the most probable path as well, and thus the corresponding word sequence, it is necessary to keep track of the argument that maximized Equation 3.7 at each time step, this is done as

$$\psi_l(j) = \arg\max_i \{\delta_{l-1}(i) a_{ij}\}. \tag{3.10}$$

Once all feature vectors have been processed, a backtracking is necessary to reconstruct the optimal path and thus obtain the optimal state sequence $\hat{\mathbf{q}} = q_1 \cdots q_L$. The optimal state in the last element of the sequence $\mathbf{q}$ obtained as

$$q_L = \arg\max_i \{\delta_{L-1}(i)\}, \tag{3.11}$$

and from it the subsequent states can be backtracked by applying

$$q_{l-1} = \psi_l(q_l). \tag{3.12}$$

A possible generalization of this implementation is the so called token passing approach [Young 1989]. In such an approach, the space of possible sequence of states is transited by a set of tokens starting at the initial states of each word. At time step l, state S_j contains zero, one or more tokens storing a particular sequence of states reaching that state $q_1 \cdots q_l = S_j$ and its corresponding joint probability $p(q_1 \cdots q_l = S_j, \mathbf{o}_1 \cdots \mathbf{o}_l)$. At the next time step, each state sends a copy of the tokens it contains to all reachable states ($S_i : a_{ji} \neq 0$). Each token is then updated to store the new state value in the sequence and update its probability count as

$$p(q_1 \cdots q_l = S_j, q_{l+1} = S_i, \mathbf{o}_1 \cdots \mathbf{o}_l, \mathbf{o}_{l+1}) = p(q_1 \cdots q_l = S_j, \mathbf{o}_1 \cdots \mathbf{o}_l) a_{ji} b_i(\mathbf{o}_{l+1}). \tag{3.13}$$

After all tokens have been propagated one time step, a rule is applied to reduce the number of tokens to a number computationally feasible to continue propagation at the next step. In the simplest case, all tokens but the one bearing the maximal probability are discarded. When the end of the utterance is reached the surviving token(s) will contain the most probable state sequence(s).

3.3 Feature Extraction

In the feature extraction process, the acoustic signal in its raw form is transformed into a domain more suitable for pattern recognition. Typically, a digitized waveform $x(t)$ is transformed into a time-feature matrix representation o_{il}, as displayed in the left part of Figure 3.1. Each vector of features is obtained by applying a feature extraction to a small segment of the time domain signal $x(t)$, typically 10ms to 30ms long, for which the process to be modeled is considered stationary. A possible example of this type of transform would be the short-time Fourier transform (STFT) seen in Chapter 2 and depicted in Figure 2.2. Given a time domain signal $x(t)$, its time-frequency representation X_{kl} can be obtained by using the STFT as

$$X_{kl} = \sum_{t'=1}^{N} x(t' + l \cdot M) h(t') \exp\left(-2\pi i \frac{(t'-1)(k-1)}{N}\right), \tag{3.14}$$

where i is the imaginary unit, $t' \in [1 \cdots N]$ is the sample index within the frame, the index $k \in [1 \cdots K]$ corresponds to the index of each Fourier bin up to half the sampling frequency and $l \in [1 \cdots L]$ is the frame index. Although the use of STFT features alone does not achieve good recognition rates, It provides a basis for some of the most successful feature extractions. The STFT domain is also the domain where speech enhancement usually takes place (see Chapter 2).
Since this domain still is sub-optimal for speech recognition, other transformations are applied to facilitate the modeling of speech, eliminate unnecessary information or enhance the relevant features. The study of speech feature extractions has been taking place for many years and remains a very active research topic. In general, a feature extraction for speech recognition involves both usual pattern recognition techniques, like decorrelation or linear prediction analysis, and techniques that aim to imitate the way humans process speech, like filterbanks that imitate the behavior of the cochlea or psycho-acoustic filters that emulate time-masking effects of the human ear and brain. The two feature extractions presented in the next sections are among the most efficient and broadly used.
Note that throughout the rest of the chapter indices i, j and k will be used to denote the following feature domains: Cepstral domain (i), filterbank domain (j) and Fourier domain (k). The same index will be used with transformations of these domains that do not alter the number of features.

Figure 3.2: Steps involved in the Mel-cepstral feature extraction. Gray colored blocks denote non-linear steps.

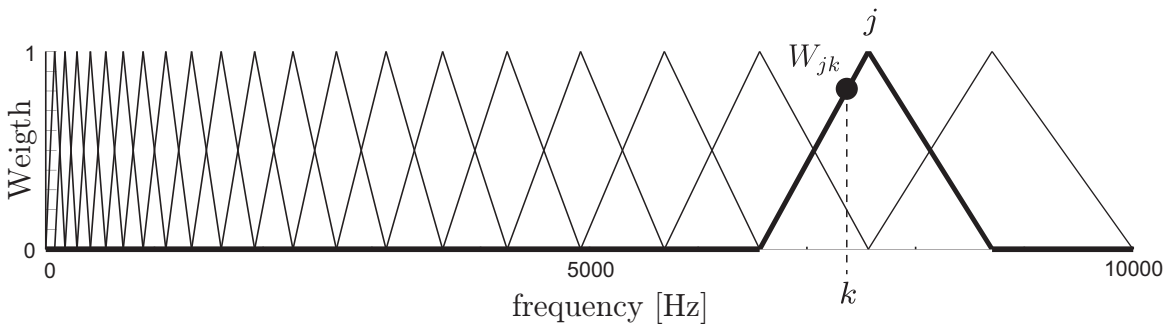

Figure 3.3: Bank of $J = 21$ Mel-spaced triangular filters for a sampling frequency of 20KHz. Weights associated with the j^{th} filterbank output are displayed in bold. The weight associated with the k^{th} frequency bin W_{jk} is shown as well.

3.3.1 Mel-Cepstral Features

Due to their high robustness and simplicity, Mel-cepstral features [Davis 1980] are widely accepted as the best option for feature extraction in ASR systems under general conditions. As seen in Figure 3.2, these features combine following transformations

1. Obtain the STFT matrix from the signal $x(t)$ as defined in Equation 3.14.
2. Extract the short-time spectral amplitude (STSA) A_{kl} of each STFT element X_{kl}.
3. Obtain each Mel-filterbank output M_{jl} from each column of the STSA, $\mathbf{A}_1 \cdots \mathbf{A}_L$.
4. Apply the logarithm to each element of the Mel-filterbank matrix M_{jl}.
5. Obtain each Mel-frequency cepstral coefficient (MFCC) C_{il} by applying the discrete cosine transform (DCT) to each column of the log-filterbank output $\log(\mathbf{M}_1) \cdots \log(\mathbf{M}_L)$[1].

The Mel-filterbank depicted in Figure 3.3 is a bank of J overlapping triangular bandpass filters equally spaced in the Mel-frequency scale f^{Mel}, which corresponds to

$$f^{\text{Mel}} = 2595 \log_{10}\left(1 + \frac{f}{700}\right). \tag{3.15}$$

where f is the conventional frequency scale.
The Mel scale was determined from human perception experiments in which subjects were asked to adjust a stimulus tone to perceptually half the pitch of a reference tone [Shannon 2003]. The Mel-filterbank is therefore able to capture the perceptually important characteristics of speech. Given a frame of STSA features, obtained as

$$A_{kl} = |X_{kl}|, \tag{3.16}$$

[1]Here, log() operates elementwise.

with $k \in [1 \cdots K]$, the j^{th} channel of the filterbank is obtained by a weighted sum of the STSAs as

$$M_{jl} = \sum_{k=1}^{K} W_{jk} A_{kl}. \tag{3.17}$$

It should also be noted that the number of features after the filterbank is reduced by around one order of magnitude with respect to the STFT domain (for typical K and J values). This is a relevant factor that will be addressed in coming chapters. To complete the Mel-cepstral features, the DCT is applied to the logarithm of the filterbank output as[2]

$$C_{il} = \sqrt{\frac{2}{J}} \sum_{j=1}^{J} \log\left(M_{jl}\right) \cos\left(\frac{\pi i}{J}(j - 0.5)\right). \tag{3.18}$$

Despite the high success of the Mel-cepstrum, this feature extraction is not completely perceptually founded. The use of the cepstral transformation is justified by principal component analysis of the features of vowels, which shows that the direction of the eigenvectors is similar to that of a cosine series expansion [Davis 1980].
Perceptually motivated feature extractions remain a very active topic nowadays, however most of the better founded models offer poorer recognition results in comparison with the Mel-cepstral features and exhibit compatibility problems with the speech enhancement algorithms[3]. This can be explained by the fact that such feature extractions are not developed exclusively for use in ASR but also for other applications like psycho-acoustical quality criteria, development of hearing aids, speech compression or in general better understanding of the human perception of speech.

3.3.2 Perceptual Linear Prediction Features

An alternative to the Mel-cepstral feature extraction which shows similar performance and has a better founded psycho-acoustical background are the Perceptual Linear Prediction features (PLP) [Hermansky 1985]. As shown in Figure 3.4, they consist of the following steps

1. Obtain the STFT matrix from the vector signal $x(t)$ as defined in Equation 3.14.
2. Extract the power spectral density (PSD) P_{kl} of each STFT element X_{kl}.
3. Obtain each Bark-filterbank output B_{jl} from each column of the PSD, $\mathbf{P}_1 \cdots \mathbf{P}_L$.
4. Multiply each column $\mathbf{B}_1 \cdots \mathbf{B}_L$ by the equal loudness pre-emphasis.
5. Raise each element to the power of 0.33 to simulate the power law of hearing and obtain each perceptually modified power spectrum coefficient H_{jl}.

[2] See [Deller 1987, pp. 364-366] for a more detailed description of the short-time cesptrum.

[3] Although not considered in the final version of this work, experiments where conducted with the psycho-acoustical model in [Tchorz 2000] as an alternative to the RASTA-PLPs. These were ultimately not used due to the above mentioned problems of compatibility.

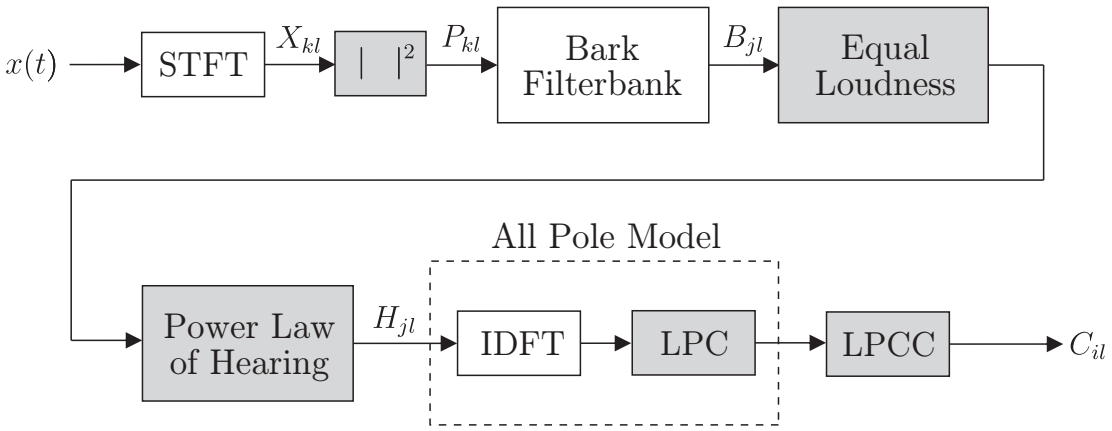

Figure 3.4: Steps involved in the PLP feature extraction with cepstra computed from the linear prediction coefficients (LPCCs and LPCs respectively). Gray colored blocks denote non-linear steps.

6. Compute an autoregressive all pole model of each column $\mathbf{H}_1 \cdots \mathbf{H}_L$ obtaining the linear prediction coefficients (LPCs).

7. Optional computing of cepstral coefficients C_{il} from the LPC (LPCCs).

The first two steps are very similar to the ones explained in the Mel-Cepstral features. The Bark filterbank is also a perceptually motivated bank of non-linearly spaced filters, with a form very similar to the one displayed in Figure 3.3, in this case corresponding to the scale.

$$f^{\text{Bark}} = 6 \ln \left(\frac{f}{600} + \sqrt{\left(\frac{f}{600} \right)^2 + 1} \right). \tag{3.19}$$

The output matrix B_{jl} of the Bark filterbank can then be computed from the PSD of the signal[4]

$$P_{kl} = |X_{kl}|^2, \tag{3.20}$$

by using

$$B_{jl} = \sum_{k=1}^{K} V_{jk} P_{kl}. \tag{3.21}$$

where V_{jk} corresponds here to the filterbank weights as in the case of the Mel-filterbank. The equal loudness pre-emphasis is an approximation of the non equal sensitivity of human hearing at different frequencies and simulates the sensitivity of hearing at about

[4]Throughout this thesis the PSD is defined as the squared magnitude of the DFT (known as periodogram). This is in fact not fully accurate since the periodogram is just an estimate of the PSD. This notation is kept for simplicity and coherence with some referred publications.

Figure 3.5: Steps involved in RASTA filtering. Gray colored blocks denote non-linear steps.

the 40 dB level. In this case, this pre-emphasis corresponds to a multiplicative gain ψ_j that only varies with filterbank index j. The approximation of the power law of hearing corresponds to a cubic root compression and simulates the non linear relation between the intensity of a sound and its perceived loudness. These two transformations together can be expressed as

$$H_{jl} = (\psi_j B_{jl})^{0.33} . \tag{3.22}$$

The final step of the PLP feature extraction is the autoregressive all-pole model of the perceptually modified power spectrum. This approximates the areas of high energy of the power spectrum, which often correspond to the resonance frequencies of the vocal tract formants. Its implementation is done via the inverse DFT and the Levinson-Durbin recursion to obtain the LPCs. Since this is a well known procedure and the techniques developed in this thesis do not require detailed knowledge about it, no explicit formulation of the algorithm will be given here. A detailed explanation of both algorithms can be found in [Deller 1987, pp. 290-302]. As an optional postprocessing step to improve feature performance, LPC based cepstral coefficients (LPCCs) can also be computed using [Davis 1980, Eq. 3]

3.3.3 RelAtive SpecTrAl Filtering

An extension of the PLP features also proposed by Hermansky is the use of RelAtive SpecTrAl (RASTA) Filtering [Hermansky 1994]. This consists of a finite impulse response (IIR) filter (see [Proakis 2007, Sec. 2.3.7]) that operates over the time dimension (index l) of the feature matrix. It is applied between steps 3 and 4 of the previously explained PLP features (see Figure 3.5) and consists of the following steps

1. Apply a non-linear compressing transformation (logarithm) to each output of step 3 of the PLP, Bark filtered PSD, B_{jl}.

2. Apply RASTA filter.

3. Return to the domain of step 3 of the PLP by applying the inverse transformation of the compression step (exponential)[5].

[5]Note that steps 4 and 5 of the PLPs (see Section 3.3.2) can also be performed in the log domain, before applying the exponential, where they become a linear transformation.

The RASTA filter emulates the preference of humans for sounds with a given rate of change, thus eliminating or attenuating artifacts that lie outside the typical rate of change of the vocal tract. The RASTA filter suppresses both slowly changing factors in speech that can be a product, for example, of convolutional noise, as well as very fast frame to frame changes resulting e.g. from speech enhancement artifacts. The filter realizes the function

$$H(z) = 0.1 * \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.94z^{-1}}, \tag{3.23}$$

or equivalently corresponds to the following difference equation

$$y_{jl} = \sum_{d=1}^{4} b_d x_{jl-d} - a_1 y_{jl-1}, \tag{3.24}$$

where y_{jl} is the l^{th} RASTA filter output, $x_{jl} \cdots x_{jl-4}$ correspond to the four previous logarithm domain inputs and y_{jl-1} to the previous RASTA filter output. The scalars $b_0 \cdots b_4, a_1$ are the normalized feedforward and feedback coefficients.

3.3.4 Pre-Emphasis

Pre-emphasis corresponds to the artificial enhancement of higher frequencies to reduce the dynamic range of the voiced speech spectra [Markel 1967]. Since this generally has a positive effect on automatic speech recognition rates, it is almost always included as part of the feature extraction. Pre-emphasis is usually applied in the time domain through the following finite impulse response (FIR) filter

$$\tilde{x}(t) = x(t) - \alpha x(t-1), \tag{3.25}$$

where α determines the cutoff frequency of the corresponding single-zero filter and usually takes values between 0.9 and 0.99. This filter has therefore a very similar effect to the equal loudness pre-emphasis applied in the frequency domain of the PLP transformation.

3.3.5 Delta and Acceleration Features

Delta and acceleration features can be added to the previous transformations to provide limited information about the time evolution of the features. Delta coefficients are computed as either a simple or weighted average of frames contiguous in time. Acceleration coefficients are then computed by applying an equivalent transformation to the delta coefficients. The delta transformation corresponds to the FIR filter

$$D_{il} = \sum_{d=-\frac{\theta}{2}}^{\frac{\theta}{2}} \Theta_d C_{il+d}, \tag{3.26}$$

where $C_{il-\frac{\theta}{2}} \cdots C_{il+\frac{\theta}{2}}$ correspond to the frames of the MFCC or LPCC cepstral coefficients over a small window of size θ and Θ_d is the corresponding weight of each frame.

3.3.6 Cepstral Mean Subtraction

Any convolutive distortion of the clean signal in the time domain, like a non ideal microphone or channel transfer function, becomes additive in the log-spectral domain. If these transfer functions are time invariant, they can be eliminated by subtracting the mean of the log-spectrum over large periods of the signal. Since the cepstral domain is only a linear transformation of the log-spectral domain, it can also be subject to this technique, which in this case is known as cepstral mean subtraction (CMS). A standard CMS procedure realizes the formula

$$\tilde{C}_{il} = C_{il} - \frac{1}{L}\sum_{l=1}^{L} C_{il}. \tag{3.27}$$

The efficiency of the CMS is limited by the adequate estimation of the channel mean. When using a conventional averaging method for the computation of the mean, large periods of the signal have to be available.

3.4 The ETSI ES 202 050 Standard for Robust Speech Recognition

The advanced front-end (AFE), defined in the standard ES 202 050 of the European Telecommunications Standards Institute (ETSI), specifies the algorithms for speech enhancement, feature extraction and transmission of the features of a distributed recognition (DSR) system. DSR systems belong to a subcategory of ASR systems in which the speech recognition task is divided between a terminal, typically a hand held device with limited computation and storage resources, and a server, which corresponds to one or more computers capable of storing and processing large amounts of data simultaneously. Excluding the decoding and transmission aspects for the communication between client and server, the advanced front-end can be considered as composed of the following steps[6], referenced in Figure 3.6.

1. The noise reduction block [Macho 2002] consists of a modified version of the two stage Mel-warped Wiener filter[7] [Agarwal 1999] with two steps of gain and noise variance estimation combining different algorithms. The output of this block is a time-domain estimation of the clean signal $\hat{x}_N(t)$, obtained by convolution of the filter impulse responses with the original noisy signal and a posterior offset compensation. Excellent suppression of residual noise is attained at this step.

2. The SNR-dependent waveform processing [Macho 2001] and pre-emphasis blocks aim to restore the periodical structure of voiced speech and enhance its higher

[6]Note that the advanced front-end includes the speech enhancement pre-processing as part of the feature extraction (e.g. It is used also during training of the ASR system). This is not always the case.

[7]Despite using the term 'Wiener filter' both gain functions and a priori SNR estimations are computed as ratios of amplitudes rather than as ratios of powers and should be thus better described as 'amplitude subtraction' [Boll 1979] related estimations.

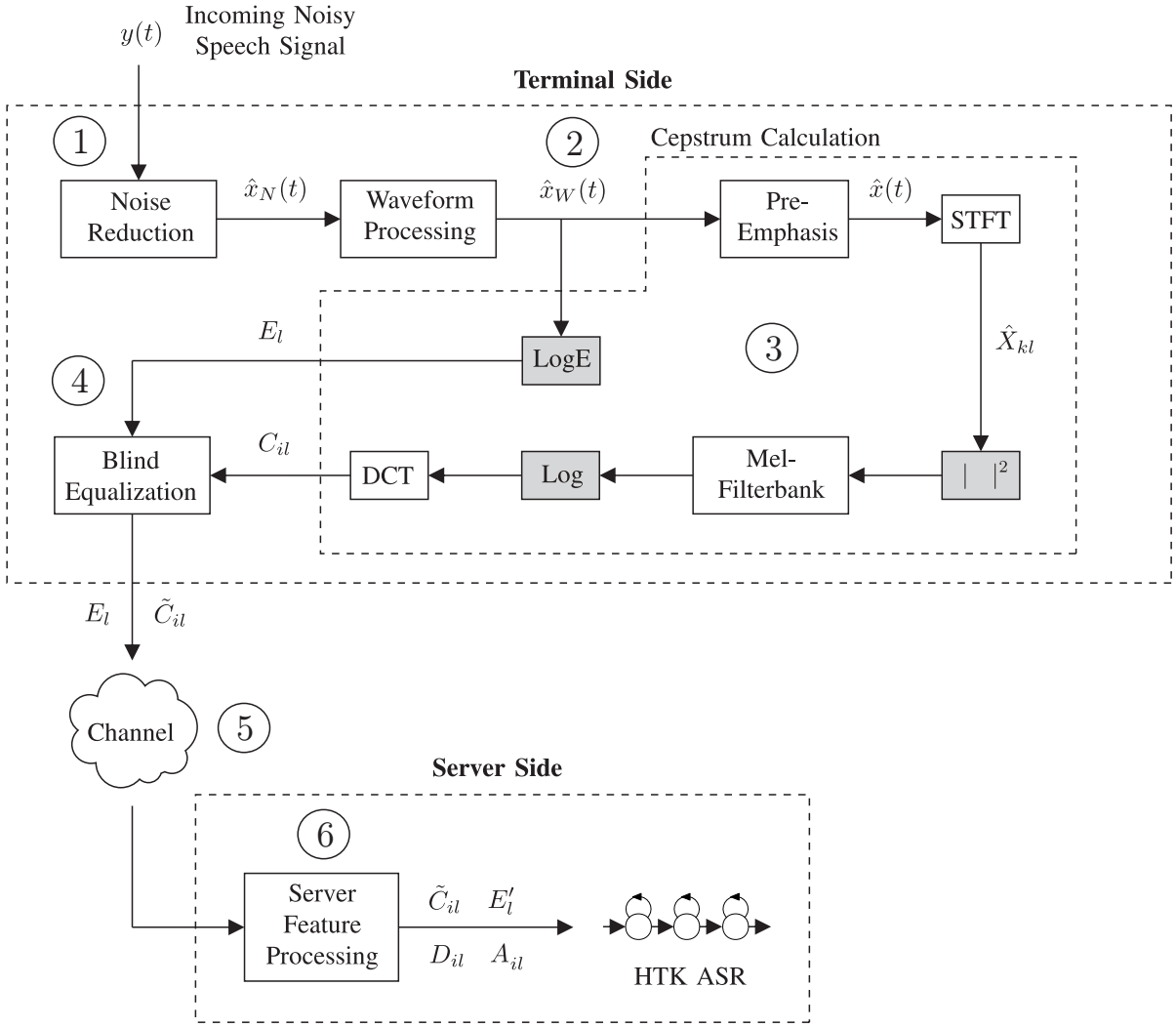

Figure 3.6: Steps involved in the ETSI ES 202 050 standard for robust speech recognition. Blocks corresponding to coding, transmission and decoding have been omitted. Gray colored blocks denote non-linear feature extraction steps.

frequency components respectively. Both steps require a high SNR to work properly. On the one hand, waveform processing depends on the identification of the speech maxima, which might fail at low SNRs. On the other hand, non-suppressed residual noises at high frequencies result in higher distortions when emphasized, which is particularly problematic in voiced segments [Vergin 1995].

3. The feature extraction in the advanced front-end corresponds to the conventional Mel-cepstral features introduced in Section 3.3.1 except for the use of PSD features instead of the STSA and the computation of additional log-energy features, described in Section 3.4.1.

4. A blind equalization of the cepstra is performed after the feature extraction to

remove remaining distortions. This is described in Section 3.4.2.

5. Previous to the transmission through the channel, a vector quantization of the signal is needed [ETSI 2007, sec. 6.6.2].

6. Server side operations include the linear combination of the first cepstral coefficient [ETSI 2007, sec. 9.1] and log-energy coefficient to obtain a new energy coefficient E'. A computation of delta and acceleration features (see Section 3.3.5) also takes place at the server side [ETSI 2007, sec. 9.2]. Finally, a second voice activity detection system is used to eliminate frames containing only noise [ETSI 2007, sec. 9.3].

3.4.1 Log-Energy Features

It is often useful to complement the cepstral coefficients with a measure of energy obtained directly from the time domain raw data. One possibility is to employ the logarithm of the energy of each frame of time domain samples as given by

$$E_l = \log\left(\sum_{t'=1}^{N} x_{t'l}^2\right), \tag{3.28}$$

where $x_{t'l}$ corresponds to the framed clean signal. This can be obtained from Equation 2.2 by replacing $y(t)$ by $x(t)$.

3.4.2 Blind Equalization

Blind equalization (BE) can be considered as a noise suppression scheme in the cepstral domain similar to cepstral mean subtraction. The particular implementation described here forms part of the ETSI standard for robust speech recognition [ETSI 2007, sec. 5.4]. This BE implementation is based on the minimization of a distortion measure between the actual signal and a given reference of the flat spectrum cepstra C_i^{ref}. Each equalized cepstral coefficient $\tilde{C}_{il}$ is obtained as

$$\tilde{C}_{il} = C_{il} - b_{il} \tag{3.29}$$

where b_{il} is a bias correction computed recursively from[8]

$$b_{il} = (1 - \mu_l) b_{il-1} + \mu_l \left(C_{il-1} - C_i^{\text{ref}}\right) \tag{3.30}$$

and where the step size factor μ_l is obtained online using the previous frame energy information (see Equation 3.28) by applying the following rule

[8] Note that the ETSI 202 050 Manual and the C-Code implementation differ on the multiplying factor of the first summand of Equation 3.30. Here, the implementation of the C-Code was used as the reference.

$$\mu_l = \begin{cases} \nu & \text{if } \frac{E_{l-1}-211}{64} >= 1 \\ \nu \cdot \frac{E_{l-1}-211}{64} & \text{if } 0 < \frac{E_{l-1}-211}{64} < 1 \\ 0 & \text{if } \frac{E_{l-1}-211}{64} =< 0, \end{cases} \tag{3.31}$$

and where $\nu = 0.0087890625$ is a fixed parameter given in [Kuroiwa 2003].

3.5 Measuring Recognition Performance

When measuring the robustness of a given ASR system, the most common method used to determine its performance is to count the number of word errors. These errors are determined by comparing the transcription output by the ASR system and a reference transcription. Three types of errors are usually differentiated:

1. Substitutions (S), which correspond to a word being mistakenly recognized as another.
2. Deletions (D), that is, omissions of a word present in the original transcription.
3. Insertions (I), which correspond to extra words not present in the original transcription.

Since there can be missing or added words in a recognized transcription, given a reference transcription and the output of the ASR system, the number of substitution, deletion and insertion errors can not be determined directly. It is first necessary to align the recognized and reference transcriptions. This is done via a dynamic programming technique in a similar fashion to the Viterbi algorithm (see Section 3.2). Once the number of each type of errors has been counted, three performance measures are usually computed from them. The word error rate (WER) accounts for the total number of errors. For a reference sentence of N words, the WER is determined as

$$\text{WER} = 100 \cdot \frac{S + D + I}{N}. \tag{3.32}$$

Along with the WER, which is the most widespread performance coefficient, word correctness (WC) and word accuracy (WA) are also often used. These are computed as

$$\text{WC} = 100 \cdot \frac{N - S - D}{N}, \tag{3.33}$$

$$\text{WA} = 100 \cdot \frac{N - S - D - I}{N}. \tag{3.34}$$

The separate performance coefficients allow to measure deletion and substitution errors separately from insertion errors. WA is the format chosen in the ETSI standard [ETSI 2007]. Note also that

$$\text{WA} = 100 - \text{WER}. \tag{3.35}$$

3.6 Automatic Speech Recognition with Uncertain Observations

In recent years, uncertainty decoding, or more generally observation uncertainty techniques, have emerged as a powerful tool to increase the accuracy of ASR systems. The main idea behind this approach is to consider the clean speech features $\mathbf{o}$ no longer deterministic but probabilistic or uncertain and ruled by a distribution $p(\mathbf{o}|I)$, where I is the available information. This uncertainty represents the lost information that causes the mismatch between the observed speech and the trained ASR model. The distribution $p(\mathbf{o}|I)$ can then be combined with the statistical models of the ASR system to yield superior speech recognition performance. Uncertainty of observation has been determined directly from speech enhancement techniques in the feature domain, like Microsoft's SPLICE [Droppo 2002] or model based feature enhancement (MBFE) [Stouten 2006]. Other models of uncertainty have been obtained by considering the relation between speech and additive noise in the logarithm domain. Using this relation, uncertainty has been computed based on a parametric model of speech distortion [Yoma 1998, Deng 2002], or the residual uncertainty after feature enhancement [Benitez 2004]. Other approaches have exploited different sources of uncertainty like channel transmission errors [Umbach 2006]. In the particular context of this thesis, the available information corresponds to the the STFT of the noisy signal $y(t)$ (see Chapter 2)[9]. This is represented by a matrix of Fourier coefficients $\mathbf{Y}$ computed using Equations 2.2 and 2.3. The probabilistic description of each frame of features $\mathbf{o}_l$ is then given by the PDF $p(\mathbf{o}_l|\mathbf{Y})$.
In order to integrate uncertain observations with the ASR system, most of these techniques employ variations of the uncertainty decoding approach [Droppo 2002]. However, modified imputation, introduced in [Kolossa 2005], has shown to provide a better performance in the context of uncertainty propagation and will therefore also be considered here.

3.6.1 Uncertainty Decoding

A common assumption to most uncertainty of observation approaches is that the uncertain description of each frame of features $\mathbf{o}_l$ is multivariate Gaussian distributed. This is also approximately true for the uncertainty propagation approach presented here, as it will be demonstrated in Chapters 6 and 7. The uncertain description of the features can then be expressed as

$$p(\mathbf{o}_l|\mathbf{Y}) \approx N\left(\mathbf{o}_l; \boldsymbol{\mu}_l^{\mathrm{o}}, \boldsymbol{\Sigma}_l^{\mathrm{o}}\right), \tag{3.36}$$

where $\boldsymbol{\mu}_l^{\mathrm{o}}$ and $\boldsymbol{\Sigma}_l^{\mathrm{o}}$ represent the mean vector and covariance matrix of the feature frame $\mathbf{o}_l$. As seen at the beginning of this chapter, in the context of hidden Markov models (HMM), the probability of being at state S_j of the HMM and generating observation

[9] The information regarding the speech enhancement could also be considered as available information here and it is ommited for simplicity

frame $\mathbf{o}_l$ is often modeled using mixtures of multivariate Gaussian distributions (see Equation 3.3) as

$$b_j(\mathbf{o}_l) = p(\mathbf{o}_l|q_l = S_j) = \sum_{m=1}^{M} W_{jm} N(\mathbf{o}_l; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}), \tag{3.37}$$

where $b_j(\mathbf{o}_l)$ is the state emission probability and the parameters W_{jm}, $\boldsymbol{\mu}_{jm}$ and $\boldsymbol{\Sigma}_{jm}$ correspond to the weight, vector mean and covariance matrix of the m^{th} Gaussian mixture in state S_j, and were previously determined during training (see Section 3.1). If instead of a deterministic description of $\mathbf{o}_l$ a probabilistic description $p(\mathbf{o}_l|\mathbf{Y})$ is available, it is possible to approximate the probability of the underlying clean features by averaging over all possible values of $\mathbf{o}_l$ as

$$\widehat{b_j(\mathbf{o}_l)} = \int b_j(\mathbf{o}_l) p\left(\mathbf{o}_l|\mathbf{Y}\right) d\mathbf{o}_l. \tag{3.38}$$

Replacing Equations 3.36 and 3.37 in this formula and taking the sum operator out of the integral results in

$$\widehat{b_j(\mathbf{o}_l)} = \sum_{m=1}^{M} W_{jm} \int N(\mathbf{o}; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}) N\left(\mathbf{o}_l; \boldsymbol{\mu}_l^{\mathrm{o}}, \boldsymbol{\Sigma}_l^{\mathrm{o}}\right) d\mathbf{o_l}. \tag{3.39}$$

where the remaining integral is solved in [Droppo 2002] and leads to

$$\widehat{b_j(\mathbf{o}_l)} = \sum_{m=1}^{M} W_{jm} N\left(\boldsymbol{\mu}_l^{\mathrm{o}}; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm} + \boldsymbol{\Sigma}_l^{\mathrm{o}}\right). \tag{3.40}$$

That is, the solution provided by uncertainty decoding is equivalent to evaluating the state emission probability using the mean of the uncertain description of $\mathbf{o}_l$ and adding the variance of this description to the state variance.
It should be noted that the original derivation of uncertainty decoding differs from the one presented here although it leads to an identical formula. Here, a probabilistic description of the clean signal is transformed through the feature extraction to yield a probabilistic description of the clean features. In [Droppo 2002] the noisy signal is directly transformed through the feature extraction instead. The uncertainty is then used to modify the state emmiting probability and dynamically learn these noisy features.
A general classification of uncertainty decoding algorithms can be found on [Liao 2008]. In this classification the method here described as uncertainty decoding is termed "front-end" uncertainty decoding. This is done to underline the difference between this an other techniques which exploit better the interaction between uncertainty and the ASR. Finally it should also be noted that the form of uncertainty decoding used here was also derived in [Yoma 2002] with a different name.

3.6.2 Modified Imputation

A possible alternative to uncertainty decoding is to find the most likely observation estimate by maximizing the joint probability of the observation given by the HMM and

the propagated uncertainty information. By assuming the uncertainty distribution to be independent of the state or mixture, this can be approximated as [Kolossa 2005]

$$\hat{\mathbf{o}}_l^j \approx \arg\max_{\mathbf{o}_l} \left\{ p(\mathbf{o}_l | q_l = S_j) p(\mathbf{o}_l | \mathbf{Y}) \right\}. \tag{3.41}$$

However it is more simple to estimate the most likely observation for each mixture m by solving

$$\begin{aligned} \hat{\mathbf{o}}_l^{jm} \approx \arg\max_{\mathbf{o}_l} \left\{ p(\mathbf{o}_l | q_l = S_j, m) p(\mathbf{o}_l | \mathbf{Y}) \right\} = \\ \arg\max_{\mathbf{o}_l} \left\{ N(\mathbf{o}_l; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}) N\left(\mathbf{o}_l; \boldsymbol{\mu}_l^{\mathrm{o}}, \boldsymbol{\Sigma}_l^{\mathrm{o}}\right) \right\}. \end{aligned} \tag{3.42}$$

where $p(\mathbf{o}_l | q_l = S_j, m)$ corresponds to the m^{th} mixture of Equation 3.37. By solving Equation 3.42 [Kolossa 2008, pp. 112-116] the following is obtained

$$\hat{\mathbf{o}}_l^{jm} = \left(\left(\boldsymbol{\Sigma}_l^{\mathrm{o}}\right)^{-1} + \boldsymbol{\Sigma}_{jm}^{-1} \right)^{-1} \boldsymbol{\mu}_{jm} \boldsymbol{\Sigma}_{jm}^{-1} + \left(\left(\boldsymbol{\Sigma}_l^{\mathrm{o}}\right)^{-1} + \boldsymbol{\Sigma}_{jm}^{-1} \right)^{-1} \boldsymbol{\mu}_l^{\mathrm{o}} \left(\boldsymbol{\Sigma}_l^{\mathrm{o}}\right)^{-1}. \tag{3.43}$$

Due to the characteristics of the cepstral coefficients the matrices $\boldsymbol{\Sigma}_{jm}$ are usually considered to be diagonal. In the tests performed in the experimental setup the covariances of the uncertainty were also considered diagonal. For such cases, Equation 3.43 simplifies to

$$\hat{o}_{il}^{jm} = \alpha_{ijm} \mu_{ijm} + (1 - \alpha_{ijm}) \mu_{il}^{\mathrm{o}}, \tag{3.44}$$

for the i^{th} feature, with

$$\alpha_{ijm} = \frac{\Sigma_{iil}^{\mathrm{o}}}{\Sigma_{iijm} + \Sigma_{iil}^{\mathrm{o}}}. \tag{3.45}$$

where $\Sigma_{iil}^{\mathrm{o}}$ and Σ_{iil} correspond to the uncertain features and model variances. Note that for values of the ratio $\Sigma_{iil}^{\mathrm{o}} / \Sigma_{iil}$ tending to infinity modified imputation converges to the missing-feature technique known as imputation [Cooke 2001]

$$\hat{o}_{il}^{jm} = \mu_{ijm}. \tag{3.46}$$

4 MMSE Speech Enhancement with GMM Priors

4.1 The Need for Supergaussian Models

As discussed in the introduction to speech enhancement in Chapter 2, a vast majority of the statistical methods for speech enhancement rely on models of speech and noise in the STFT domain under three basic assumptions

1. Speech and noise are instationary, non-ergodic random processes.
2. Speech is corrupted by additive noise.
3. Speech and noise are modeled as zero mean Gaussian processes.

Although the last two assumptions have been put into question often, this chapter will be centered only on the third, which has the most relevant implications for the algorithms and techniques developed in following chapters.
In order to proceed with an analysis of the Gaussian assumption, it is first necessary to be more specific in its definition. As seen in Chapter 2 and Figure 2.2, the STFT of a signal

$$y(t) = x(t) + d(t), \tag{4.1}$$

containing speech $x(t)$ and noise $d(t)$, corresponds to a matrix of complex valued elements Y_{kl} containing a time-frequency representation of $y(t)$. The index $k \in [1 \cdots K]$ corresponds to the index of each Fourier bin up to half the sampling frequency and $l \in [1 \cdots L]$ is the frame index. Due to the assumption of instationarity and non-ergodicity each time-frequency element is considered as being ruled by its own random process, independent of the others. By considering also the assumption of additiveness of noise, each Fourier coefficient Y_{kl} can be seen as generated by the interaction of two statistically independent hidden processes, the Fourier coefficient of the clean signal X_{kl} and the Fourier coefficient of the noise signal D_{kl}, as

$$Y_{kl} = X_{kl} + D_{kl}. \tag{4.2}$$

By assuming different probability distributions for X_{kl} and D_{kl}, prior information is introduced into the model, which together with the interaction rule given in Equation 4.2 allows us to compute the likelihood and a posteriori distributions from which the clean signal can be estimated (see derivations throughout Section 2.1). An a priori model should summarize what is known about the hidden processes. In theory, such information

could be obtained directly by observing multiple samples of clean speech and noise and computing, for example, its histograms. However, since signals are not stationary and non-ergodic, different time-frequency elements of the STFT of the signal correspond to different statistical processes [Ephraim 1984] and thus it is not possible to deduce probability distributions by using histograms of the long-term behavior of speech. The only piece of a priori information available is, in principle, the fact that each Fourier coefficient is obtained as a weighted sum of random variables (see Equation 2.3). As commented in Section 2.1.1, due to the central limit theorem this justifies the assumption of Gaussianity for a frame of infinite length [Pearlman 1978, Th. 1]. This, together with the fact that estimators based on Gaussian models are easy to compute, has led to widespread use of complex Gaussian models for speech enhancement applications [McAulay 1980, Ephraim 1984, Ephraim 1985]. The use of non-Gaussian speech models, like the approach in [Porter 1984], remained marginal until recently.

4.1.1 Prior Distributions Conditioned on the a Priori SNR

An approach which allows the computation of the prior distributions of speech and noise from available data was presented by Martin in [Martin 2002]. This approach is based on the fact that many speech enhancement methods, like for example the Wiener filter presented in Section 2.1.4,

$$\hat{X}^{\mathrm{W}} = \frac{\xi}{1+\xi} \cdot Y = G(\xi)^{\mathrm{W}} \cdot Y, \tag{4.3}$$

depend on the estimation of the a priori SNR ξ. Given this dependence, Martin stated that the optimal estimator, and therefore its prior distributions, should be characterized

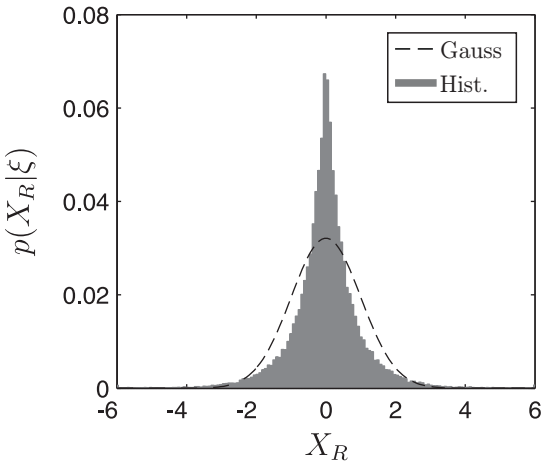

Figure 4.1: Distribution of the real component of a speech Fourier coefficient X_R (shadowed) approximated by a histogram of speech samples obtained for a narrow interval of ξ values of 19dB-21dB. Gaussian distribution plotted for comparison (dashed).

with respect to this parameter.
Using this assumption the conventional prior for the real component of the clean speech Fourier coefficient $p(X_R)$ was computed as the distribution of X_R conditioned on a narrow interval $[\xi_{\min}, \xi_{\max}]$ of ξ values as given by

$$X_R \rightarrow p(X_R|\xi \in [\xi_{\min}, \xi_{\max}]). \tag{4.4}$$

Since for this conditional distribution there are multiple samples available, the computation of histograms is straightforward. An example of the histogram of the prior $p(X_R|\xi)$ is displayed in Figure 4.1.
Based on this approach, supergaussian a priori distributions for speech, like those employing Laplace or Gamma [Martin 2002, Lotter 2005], generalized Gamma [Erkelens 2007b] or Rayleigh mixture model (RMM) [Erkelens 2007] distributions, have been demonstrated to offer a better fit to the empirically obtained distribution of the speech Fourier coefficients. MAP [Lotter 2005] and MMSE [Martin 2002, Erkelens 2007b] estimators derived under these models have also shown to yield better speech enhancement properties in terms of SNR.

4.1.2 Prior Distributions Conditioned on Knowledge of the Variance

The use of supergaussian prior distributions was also addressed by Ephraim and Cohen in [Ephraim 2004] by proposing a generative model for the speech prior, which would justify the use of both Gaussian and supergaussian priors. Ephraim stated that the form of the prior distribution of the real or imaginary components of each Fourier coefficient was dependent on knowledge of the variance σ_X^2 of each prior. More precisely, if we consider the real component of a speech Fourier coefficient X_R, Ephraim and Cohen assumed the distribution of speech conditioned by the knowledge of its variance $p(X_R|\sigma_X^2)$ to be Gaussian. The distribution of the variance $p(\sigma_X^2)$ was then used to summarize the prior knowledge regarding the variance σ_X^2. As a consequence of this, for a given assumption about the distribution $p(\sigma_X^2)$, the prior for the real component of the Fourier coefficient X_R can be computed using the following marginalization

$$p(X_R) = \int p(X_R, \sigma_X^2) d\sigma_X^2 = \int p(X_R|\sigma_X^2) p(\sigma_X^2) d\sigma_X^2. \tag{4.5}$$

The resulting expression is a continuous Gaussian mixture model, where the mixture weight distribution corresponds to the distribution of the speech variance, and thus justifies multiple priors. For example, setting $p(\sigma_X^2)$ equal to the exponential distribution in Equation 4.5 leads to the Laplace distribution used in [Martin 2002][1].

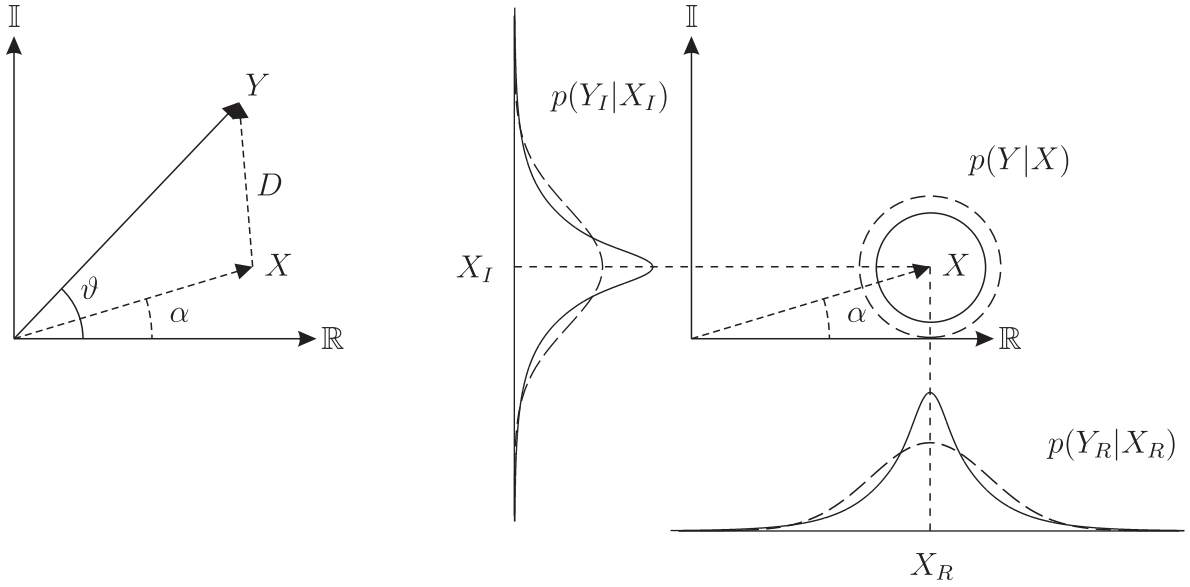

Figure 4.2: **Left:** Additive model for the noise Fourier coefficient. The observable noisy Fourier coefficient Y is obtained as a sum of the unknown clean signal Fourier coefficient X and the noise Fourier coefficient D. **Right:** Likelihood function, comparison between Gaussian (dashed) and Gaussian-mixture (continuous) marginal distributions is shown here.

4.2 The Complex Gaussian Mixture Model

4.2.1 Definition

Let X, $D \in \mathbb{C}$ be two complex random variables representing a STFT Fourier coefficient of the speech and noise processes respectively. Given the Gaussian model defined by Equations 2.11, 2.12 and 2.15 (see Chapter 2 for a detailed description), let us redefine the speech and noise prior distributions, $p(X)$ and $p(D)$, as the following mixture of zero mean circularly symmetric complex Gaussian distributions

$$p(X) = \sum_{i=1}^{I} w_{X_i} \frac{1}{\pi \lambda_{X_i}} \exp\left(-\frac{|X|^2}{\lambda_{X_i}}\right), \tag{4.6}$$

$$p(D) = \sum_{j=1}^{J} w_{D_j} \frac{1}{\pi \lambda_{D_j}} \exp\left(-\frac{|D|^2}{\lambda_{D_j}}\right), \tag{4.7}$$

where the weights are normalized as

$$\sum_{i=1}^{I} w_{X_i} = 1, \tag{4.8}$$

[1]Note that here it is implicitly assumed that the distribution of the real and imaginary components $p(X_R)$, $p(X_I)$ defines the prior $p(X)$. This does not hold in cases in which $p(X) \neq p(X_R)p(X_I)$.

$$\sum_{j=1}^{J} w_{D_j} = 1. \tag{4.9}$$

Considering the condition of additiveness in Equation 2.15, the likelihood distribution can be assumed to be centered in X and with a distribution equal to that of Equation 4.7

$$p(Y|X) = \sum_{j=1}^{J} w_{D_j} \frac{1}{\pi\lambda_{D_j}} \exp\left(-\frac{|Y-X|^2}{\lambda_{D_j}}\right). \tag{4.10}$$

The model described by Equations 4.6, 4.7 and 4.2 defines a complex Gaussian mixture model[2] under the assumption of additive noise. This model can be considered an extension of the model presented in Chapter 2 (See Figure 4.2), where the third basic assumption has been relaxed to allow this particular kind of non-Gaussian models. As it will be demonstrated, this statistical model is also an extension of the model used to derive the RMM-MMSE estimator proposed in [Erkelens 2007].

4.2.2 Properties

Since the models in Equations 4.6 and 4.7 are just weighted sums of Equations 2.11 and 2.12, the cartesian (real and imaginary) and polar (amplitude and phase) marginal distributions can be easily obtained by marginalization in a similar way to how it was done in Section 2.3.2. If we consider the speech prior distribution given by Equation 4.6, we have that the corresponding distributions for the real and imaginary components yield

$$p(X_R) = \int p(X)dX_I = \int p(X_R, X_I)dX_I = \sum_{i=1}^{I} w_{X_i} \frac{1}{\sqrt{2\pi\sigma_{X_i}^2}} \exp\left(-\frac{X_R^2}{2\sigma_{X_i}^2}\right), \tag{4.11}$$

$$p(X_I) = \int p(X)dX_R = \int p(X_R, X_I)dX_R = \sum_{i=1}^{I} w_{X_i} \frac{1}{\sqrt{2\pi\sigma_{X_i}^2}} \exp\left(-\frac{X_I^2}{2\sigma_{X_i}^2}\right), \tag{4.12}$$

where $\lambda_{X_i} = 2\sigma_{X_i}^2$. With respect to the original distributions of the complex Gaussian model defined in Equations 2.7 and 2.11, the marginal distributions of this new model are Gaussian mixture models (GMM) of the original marginals with same weights as in Equation 4.6. However, the joint distribution in Equation 4.6 does not factorize in this case into the product of the marginals in Equations 4.11 and 4.12

$$p(X) = p(X_R, X_I) \neq p(X_R)p(X_I). \tag{4.13}$$

[2]This will be referred as 'Gaussian mixture model' for speech enhancement in analogous form to the 'Gaussian model' introduced in Chapter 2.

Consequently, real and imaginary components can not be considered statistically independent under this model. This is not the case for the polar marginals, since the amplitude distribution

$$p(A) = \int_0^{2\pi} p(A,\alpha)d\alpha = \sum_{i=1}^{I} w_{X_i} \frac{2A}{\lambda_{X_i}} \exp\left(-\frac{A^2}{\lambda_{X_i}}\right), \tag{4.14}$$

results in a Rayleigh mixture model, whereas the phase remains uniformly distributed

$$\begin{aligned} p(\alpha) = \int_0^{\infty} \sum_{i=1}^{I} w_{D_j} \frac{A}{\pi\lambda_{X_i}} \exp\left(-\frac{|Ae^{i\alpha}|^2}{\lambda_{X_i}}\right) dA = \\ \frac{1}{2\pi} \int_0^{\infty} \sum_{i=1}^{I} w_{D_j} \frac{2A}{\lambda_{X_i}} \exp\left(-\frac{A^2}{\lambda_{X_i}}\right) dA = \frac{1}{2\pi}. \end{aligned} \tag{4.15}$$

Since Equation 4.6 factorizes into Equations 4.14 and 4.15, the amplitude and phase can be considered to be statistically independent as in the conventional Gaussian model.

$$p(X) = p(A,\alpha) = p(A)p(\alpha). \tag{4.16}$$

4.3 Complex Gaussian Mixture Model Based Estimators

4.3.1 Differences with Respect to the Gaussian Model

Due to the similarities between the complex Gaussian and the complex Gaussian mixture models, solving the integrals of the MMSE estimators seen in Section 2.1 bears little additional difficulty. One of the main differences between both models is the dependency between real and imaginary components of the prior and the likelihood functions which invalidates, in principle, the use of the Wiener filter approach (see Section 2.1.4). The approach used in the Ephraim-Malah filters remains however valid (see Section 2.1.5). In [Erkelens 2007] MMSE-STSA and MMSE-LSA estimators were derived under the assumption of a Rayleigh mixture model amplitude prior as in Equation 4.14 and a Gaussian distributed likelihood as in Equation 2.16. The approach presented here is a generalization of this idea, in which both prior and likelihood function are described by Gaussian mixture models as in Equations 4.6 and 4.10. In order to be able to derive MMSE amplitude estimators for these models, it is first necessary to marginalize the amplitude from the likelihood distribution as it was done with the conventional Gaussian model in Section 2.3.6. Since amplitude and phase of the clean Fourier coefficient remain statistically independent (Equation 4.16), this can be done as

$$p(Y|A) = \int_0^{2\pi} p(Y,\alpha|A)d\alpha = \int_0^{2\pi} p(Y|A,\alpha)p(\alpha)d\alpha = \frac{1}{2\pi}\int_0^{2\pi} p(Y|A,\alpha)d\alpha. \tag{4.17}$$

With exception of the summation operand and the weights, which do not affect the integral, this formula was already solved in Section 2.3.6 and leads to

$$p(Y|A) = \sum_{j=1}^{J} w_{D_j} \frac{1}{\pi\lambda_{D_j}} \exp\left(-\frac{|Y|^2 + A^2}{\lambda_{D_j}}\right) I_0\left(\frac{2|Y|A}{\lambda_{D_j}}\right). \tag{4.18}$$

4.3.2 Rice Mixture Posterior Distribution

As it was done with the Gaussian model in Section 2.1.5, the likelihood function for the complex Gaussian mixture model $p(Y|X)$ in Equation 4.18 together with the corresponding prior $p(X)$ in Equation 4.14 can be used to compute the solution to the MMSE-STSA and MMSE-LSA estimators by solving Equations 2.36 and 2.46. However, with aim in the mathematical derivations of Chapters 5 and 6, the complete posterior distribution $p(A|Y)$ will be here derived instead. Obtaining the MMSE-STSA and MMSE-LSA estimators from $p(A|Y)$ is a trivial operation explained in Section 4.3.4.
The posterior distribution $p(A|Y)$ can be obtained by applying Bayes theorem as

$$p(A|Y) = \frac{p(Y|A)p(A)}{p(Y)}. \tag{4.19}$$

This can be solved by first computing the denominator as

$$p(Y) = \int_0^\infty p(Y|A)p(A)dA =$$
$$\int_0^\infty \sum_{j=1}^{J} w_{D_j} \frac{1}{\pi\lambda_{D_j}} \exp\left(-\frac{|Y|^2 + A^2}{\lambda_{D_j}}\right) I_0\left(\frac{2|Y|A}{\lambda_{D_j}}\right) \sum_{i=1}^{I} w_{X_i} \frac{2A}{\lambda_{X_i}} \exp\left(-\frac{A^2}{\lambda_{X_i}}\right) dA, \tag{4.20}$$

where after applying the distributive property and taking sum operators out of the integral we obtain

$$p(Y) =$$
$$\sum_{i=1}^{I}\sum_{j=1}^{J} w_{D_j} w_{X_i} \int_0^\infty \frac{1}{\pi\lambda_{D_j}} \exp\left(-\frac{|Y|^2 + A^2}{\lambda_{D_j}}\right) I_0\left(\frac{2|Y|A}{\lambda_{D_j}}\right) \frac{2A}{\lambda_{X_i}} \exp\left(-\frac{A^2}{\lambda_{X_i}}\right) dA. \tag{4.21}$$

Since the only difference between the distributions obtained for the Gaussian and the Gaussian mixture models are the weighted sums, the term remaining inside of the integral is equal to the product of the prior and likelihood distributions given in Equations 2.34 and 2.33 for the Gaussian case. Consequently, integrating the amplitude out of the product of the likelihood and prior distributions leads to the distribution of the noisy Fourier coefficient given in Equation 2.18[3]. The denominator of Equation 4.19 is then equal to

$$p(Y) = \sum_{i=1}^{I}\sum_{j=1}^{J} w_{D_j} w_{X_i} \frac{1}{\pi(\lambda_{X_i} + \lambda_{D_j})} \exp\left(-\frac{|Y|^2}{\lambda_{X_i} + \lambda_{D_j}}\right). \tag{4.22}$$

[3]Let $p_G(A)$, $p_G(Y|A)$, $p_G(Y)$ define the prior, likelihood and noisy Fourier coefficient distributions in the Gaussian case. It is easy to see that $\int_0^\infty p_G(Y|A)p_G(A)dA = \int_0^\infty p_G(Y,A)dA = p_G(Y)$.

Regarding the numerator of Equation 4.19 we have

$$p(Y|A)p(A) = \sum_{j=1}^{J} w_{D_j} \frac{1}{\pi\lambda_{D_j}} \exp\left(-\frac{|Y|^2 + A^2}{\lambda_{D_j}}\right) I_0\left(\frac{2|Y|A}{\lambda_{D_j}}\right) \sum_{i=1}^{I} w_{X_i} \frac{2A}{\lambda_{X_i}} \exp\left(-\frac{A^2}{\lambda_{X_i}}\right), \quad (4.23)$$

which can be expressed as

$$p(Y|A)p(A) = \sum_{i=1}^{I}\sum_{j=1}^{J} w_{D_j} w_{X_i} \frac{2A}{\pi\lambda_{D_j}\lambda_{X_i}} \exp\left(-\frac{|Y|^2 + A^2}{\lambda_{D_j}} - \frac{A^2}{\lambda_{X_i}}\right) I_0\left(\frac{2|Y|A}{\lambda_{D_j}}\right). \quad (4.24)$$

The fractions of the exponential can then be re-organized as

$$\begin{aligned} -\frac{|Y|^2 + A^2}{\lambda_{D_j}} - \frac{A^2}{\lambda_{X_i}} + \frac{|Y|^2}{\lambda_{X_i} + \lambda_{D_j}} - \frac{|Y|^2}{\lambda_{X_i} + \lambda_{D_j}} = \\ -\frac{\lambda_{X_i} + \lambda_{D_j}}{\lambda_{X_i}\lambda_{D_j}} \left(\frac{\lambda_{X_i}|Y|^2}{\lambda_{X_i} + \lambda_{D_j}} + \frac{\lambda_{X_i} + \lambda_{D_j}}{\lambda_{X_i} + \lambda_{D_j}} A^2 - \frac{\lambda_{X_i}\lambda_{D_j}|Y|^2}{\left(\lambda_{X_i} + \lambda_{D_j}\right)^2}\right) - \frac{|Y|^2}{\lambda_{X_i} + \lambda_{D_j}} = \\ -\frac{\lambda_{X_i} + \lambda_{D_j}}{\lambda_{X_i}\lambda_{D_j}} \left(\frac{\lambda_{X_i}^2}{\left(\lambda_{X_i} + \lambda_{D_j}\right)^2}|Y|^2 + A^2\right) - \frac{|Y|^2}{\lambda_{X_i} + \lambda_{D_j}}. \end{aligned} \quad (4.25)$$

Replacing this term into Equation 4.24 we obtain

$$p(Y|A)p(A) = \sum_{i=1}^{I}\sum_{j=1}^{J} \tilde{\Omega}_{ij} \frac{2A}{\pi\lambda_{ij}} \exp\left(-\frac{\left|\frac{\lambda_{X_i}}{\lambda_{X_i}+\lambda_{D_j}} Y\right|^2 + A^2}{\lambda_{ij}}\right) I_0\left(\frac{2|Y|A}{\lambda_{D_j}}\right), \quad (4.26)$$

where

$$\tilde{\Omega}_{ij} = w_{D_j} w_{X_i} \frac{1}{\lambda_{X_i} + \lambda_{D_j}} \exp\left(-\frac{|Y|^2}{\lambda_{X_i} + \lambda_{D_j}}\right). \quad (4.27)$$

and λ_{ij} is defined equivalently to [Ephraim 1984, Eq. A.3] as

$$\lambda_{ij} = \frac{\lambda_{X_i}\lambda_{D_j}}{\lambda_{X_i} + \lambda_{D_j}}, \quad (4.28)$$

By replacing the denominator $p(Y)$ in Equation 4.22 and the numerator $p(Y|A)p(A)$ in Equation 4.26 into the Bayes theorem formula in Equation 4.19, the final form of the posterior of the complex Gaussian mixture model can be obtained as the following mixture of Rice distributions

$$p(A|Y) = \sum_{i=1}^{I}\sum_{j=1}^{J} \Omega_{ij} \frac{2A}{\lambda_{ij}} \exp\left(-\frac{\left|\frac{\lambda_{X_i}}{\lambda_{X_i}+\lambda_{D_j}} Y\right|^2 + A^2}{\lambda_{ij}}\right) I_0\left(\frac{2\left|\frac{\lambda_{X_i}}{\lambda_{X_i}+\lambda_{D_j}} Y\right| A}{\lambda_{ij}}\right), \quad (4.29)$$

with Ω_{ij} being the normalized weight

$$\Omega_{ij} = \frac{\tilde{\Omega}_{ij}}{\sum_{i=1}^{I}\sum_{j=1}^{J}\tilde{\Omega}_{ij}}. \tag{4.30}$$

The weights w_{X_i}, w_{D_j} of the original mixture models in Equations 4.6 and 4.7 can be interpreted as the discrete probability distributions of each mixture

$$p(i) = w_{X_i}, \tag{4.31}$$

and

$$p(j) = w_{D_j}, \tag{4.32}$$

where $p(X|i)$ and $p(D|j)$ would correspond to each Gaussian mixture of Equations 4.6 and 4.7 respectively and $p(Y|i,j)$ to each mixture of Equation 4.22. Given this interpretation, it is easy to see that from Equation 4.30

$$\Omega_{ij} = \frac{p(i)p(j)p(Y|i,j)}{\sum_{i=1}^{I}\sum_{j=1}^{J}p(i)p(j)p(Y|i,j)} = \frac{p(Y,i,j)}{p(Y)} = p(i,j|Y). \tag{4.33}$$

That is, the weights of the posterior correspond to the probability of the $(i,j)^{th}$ mixture given the observable information Y and $p(A|Y,i,j)$ corresponds to the $(i,j)^{th}$ Rice distributed mixture of Equation 4.29.

4.3.3 Rice Posterior Distribution

The posterior derived in the previous section is a generalization of the posterior that would be obtained for the conventional Gaussian model introduced in Chapter 2. Given the priors of the Gaussian mixture model defined in Equations 4.6 and 4.7, if we set the number of mixtures to one ($I = 1$, $J = 1$), we obtain the priors of the Gaussian model in Equations 2.11 and 2.12. In this case, the posterior distribution in Equation 4.29 yields

$$p(A|Y)\Big|_{I=1,J=1} = \frac{2A}{\lambda}\exp\left(-\frac{\left|\frac{\lambda_X}{\lambda_X+\lambda_D}Y\right|^2 + A^2}{\lambda}\right) I_0\left(\frac{2\left|\frac{\lambda_X}{\lambda_X+\lambda_D}Y\right|A}{\lambda}\right). \tag{4.34}$$

That is, for the conventional Gaussian model, the posterior distribution of the amplitude of the clean Fourier coefficient A, given the noisy Fourier coefficient Y is a Rice distribution. Since MMSE estimators require the computation of a posterior expectation $E\{A|Y\}$ (see Sections 2.3.6 and 2.3.7), the MMSE-STSA and MMSE-LSA estimators can be directly computed from the posterior $p(A|Y)$.
Furthermore, extending the MMSE-STSA and MMSE-LSA estimators to the Gaussian mixture model case is trivial, since the only difference between the posterior distributions of the Gaussian and Gaussian mixture models are the weighted sums, which do not affect the computation of the integrals.

4.3.4 MMSE Amplitude Estimators

The developed estimators will be termed Rice mixture model minimum mean square error (RiMM-MMSE) estimators.
Given the results at the end of Section 4.3.2 the posterior $p(A|Y)$ in Equation 4.29 can be written as

$$p(A|Y) = \sum_{i=1}^{I}\sum_{j=1}^{J} \Omega_{ij} p(A|Y,i,j), \tag{4.35}$$

where $p(A|Y,i,j)$ would correspond to the $(i,j)^{th}$ Rice distributed mixture of Equation 4.29.
In accordance with Equation 2.107, the RiMM-MMSE-STSA estimator would correspond to the solution of

$$\hat{A}^{\text{RiMM-STSA}} = E\{A|Y\} = \int A \frac{p(Y|A)p(A)}{p(A)} dA \int A p(A|Y) dA, \tag{4.36}$$

which can be expressed in terms of the Rice mixtures as

$$\hat{A}^{\text{RiMM-STSA}} = \int A \sum_{i=1}^{I}\sum_{j=1}^{J} \Omega_{ij} p(A|Y,i,j) dA = \sum_{i=1}^{I}\sum_{j=1}^{J} \Omega_{ij} E\{A|Y,i,j\}. \tag{4.37}$$

According to what was seen in Section 4.3.3, $E\{A|Y,i,j\}$ is the posterior expectation obtained under the conventional Gaussian model and is given by Equation 2.37. Taking the noisy signal amplitude $|Y|$ as a common factor in each of the summands $E\{A|Y,i,j\}$, we obtain the following gain function

$$G(\xi,\gamma)^{\text{RiMM-STSA}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \Omega_{ij} G(\xi_{ij},\gamma_j)^{\text{STSA}}, \tag{4.38}$$

where $G(\xi_{ij},\gamma_j)^{\text{STSA}}$ is given in Equation 2.41 and ξ_{ij} and γ_j correspond to the a priori and a posteriori signal to noise ratios for the $(i,j)^{th}$ Rice mixture

$$\xi_{ij} = \frac{\lambda_{X_i}}{\lambda_{D_j}}, \tag{4.39}$$

$$\gamma_j = \frac{|Y|^2}{\lambda_{D_j}}. \tag{4.40}$$

The same approach can be applied to compute the RiMM-MMSE-LSA estimator. This is obtained as the solution to Equation 2.46 as

$$\hat{A}^{\text{RiMM-LSA}} = \exp\left(\frac{d}{d\mu} E\{A^\mu|Y\}\Big|_{\mu=0} \right). \tag{4.41}$$

In this case, since the weights do not depend on μ, the weighted sum can be taken out of the derivative as

$$\hat{A}^{\text{RiMM-LSA}} = \exp\left(\sum_{i=1}^{I}\sum_{j=1}^{J}\Omega_{ij}\frac{d}{d\mu}E\{A^{\mu}|Y,i,j\}\Big|_{\mu=0}\right). \tag{4.42}$$

A weighted arithmetical average is transformed into a weighted geometrical average by the exponential giving

$$\hat{A}^{\text{RiMM-LSA}} = \prod_{i=1}^{I}\prod_{j=1}^{J}\exp\left(\frac{d}{d\mu}E\{A^{\mu}|Y,i,j\}\Big|_{\mu=0}\right)^{\Omega_{ij}}. \tag{4.43}$$

Taking into account the fact that the solution to $\exp\left(\frac{d}{d\mu}E_{ij}\{A^{\mu}|Y\}|_{\mu=0}\right)$ is given by Equation 2.127 and that

$$\prod_{i=1}^{I}\prod_{j=1}^{J}|Y|^{\Omega_{ij}} = |Y|^{\sum_{i=1}^{I}\sum_{j=1}^{J}\Omega_{ij}} = |Y|, \tag{4.44}$$

we obtain the final expression of the RiMM-MMSE-LSA gain as

$$G(\xi,\gamma)^{\text{RiMM-LSA}} = \prod_{i=1}^{I}\prod_{j=1}^{J}\left(G(\xi_{ij},\gamma_j)^{\text{LSA}}\right)^{\Omega_{ij}}. \tag{4.45}$$

Note that in order to compute the gain functions, the phase estimator of the complex Gaussian in Equation 2.40 was used. No particular estimator for the phase under the Gaussian mixture model was derived.

4.3.5 Estimating the Statistics of Each Mixture

In Section 2.2 various algorithms were introduced for the estimation of the statistical parameters of the clean and noise Fourier coefficients. Under the assumption of a zero mean Gaussian model, the knowledge of the variances λ_X, λ_D was sufficient to define the model. In case of the Gaussian mixtures it is necessary to estimate the variances of each mixture λ_{X_i}, $i \in [1\cdots I]$, λ_{D_j}, $j \in [1\cdots J]$. These can be computed from the total variances by noting that by the definition of zero mean variance of a complex random variable we have

$$\begin{aligned}\lambda_X = E\{|X|^2\} = \int |X|^2 \sum_{i=1}^{I} w_{X_i}\frac{1}{\pi\lambda_{X_i}}\exp\left(-\frac{|X|^2}{\lambda_{X_i}}\right)dX = \\ \sum_{i=1}^{I} w_{X_i}\int |X|^2\frac{1}{\pi\lambda_{X_i}}\exp\left(-\frac{|X|^2}{\lambda_{X_i}}\right)dX = \sum_{i=1}^{I} w_{X_i}\lambda_{X_i},\end{aligned} \tag{4.46}$$

and consequently

$$\lambda_D = \sum_{j=1}^{J} w_{D_j}\lambda_{D_j}. \tag{4.47}$$

If we now define the ratios between the total variance and each mixture variance as

$$r_{X_i} = \frac{\lambda_{X_i}}{\lambda_X}, \tag{4.48}$$

$$r_{D_j} = \frac{\lambda_{D_j}}{\lambda_D}, \tag{4.49}$$

the a priori and a posteriori signal to noise ratios of each mixture can then be expressed as

$$\xi_{ij} = \frac{r_{X_i}\lambda_X}{r_{D_j}\lambda_D}, \tag{4.50}$$

$$\gamma_j = \frac{|Y|^2}{r_{D_j}\lambda_D}. \tag{4.51}$$

4.3.6 Relation to the Rayleigh Mixture Model Estimator

As commented at the beginning of this chapter, the RMM-MMSE estimator proposed in [Erkelens 2007] is a particular case of the RiMM-MMSE estimator. The Rayleigh mixture speech prior employed in [Erkelens 2007] is equivalent to setting $J = 1$ in Equation 4.7 of the Gaussian mixture model proposed in this chapter.
To obtain the mixture weights of the RMM-MMSE estimator, we set $J = 1$ in Equation 4.27 and after multiplying by constant factors that would be removed by the normalization, the unnormalized weight yields

$$\tilde{\Omega}_i = \frac{w_{X_i}}{1+\xi_i} \exp\left(\frac{\xi_i}{1+\xi_i}\gamma\right), \tag{4.52}$$

which corresponds to the weight function g_j given in [Erkelens 2007].

4.4 Noise Prior conditioned by Noise Variance Estimation

In order to use the RiMM-MMSE-STSA and RiMM-MMSE-LSA estimators presented in the previous section, it is necessary to develop methods that allow for the estimation not only of the conventional noise and speech statistics λ_D and λ_X but also the parameters r_{D_j}, w_{D_j}, r_{X_i} and w_{X_i}. A possibility is to use the method proposed by Martin (see Equation 4.4), to determine the parameters of the marginal distribution $p(X_R)$ (see Equation 4.11) by using the well known EM algorithm for Gaussian mixture models [Press 2007, Sec. 16.1][4].
A new method is introduced here to train the noise prior. This method uses the fact that, unlike in the case of each noise Fourier coefficient D_{kl}, the variance of noise $\lambda_{D_{kl}} = 2\sigma_D^2$

[4] This is similar to the method used in [Erkelens 2007] for the estimation of the Rayleigh-mixture model parameters.

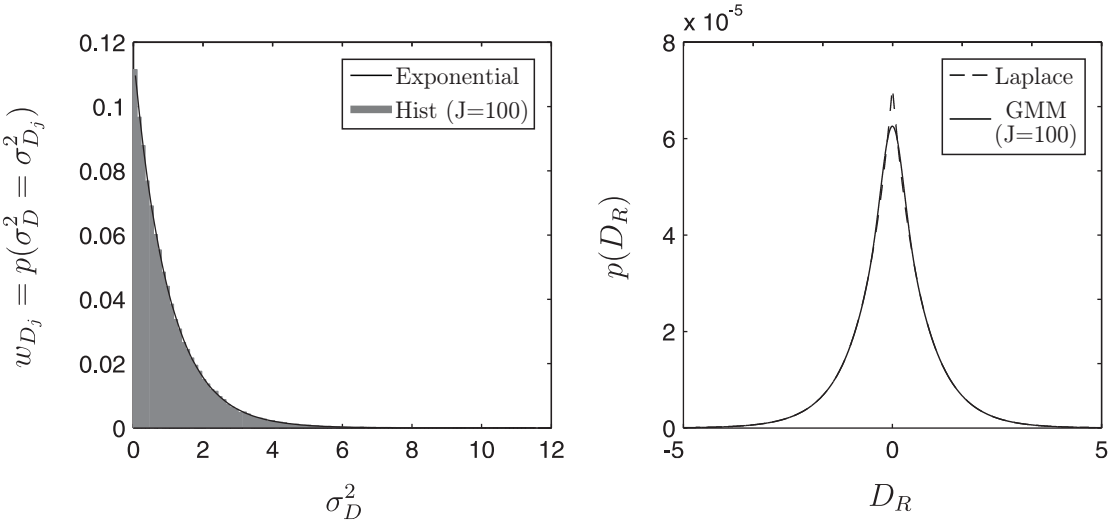

Figure 4.3: **Left:** Exponential distributed variance $p(\sigma_D^2)$ (solid). Approximation of this distribution with a $J = 100$ bin histogram $p(\sigma_D^2 = \sigma_{D_j}^2)$ (shaded). **Right:** Laplace distributed prior of D_R (dashed) and GMM prior from the histogram on the left (solid).

can be regarded as an ergodic process. This implies that for noise Fourier coefficients of different frequencies and frames, their variance $\sigma_{D_{kl}}^2$ can be considered as belonging to the same statistical process and ruled by a single distribution $p(\sigma_D^2)$. Furthermore if we consider the continuous Gaussian mixture model proposed by Ephraim and Cohen (see Equation 4.5) we have

$$p(D_R; \sigma_D^2) = \int p(D_R|\sigma_D^2)p(\sigma_D^2)d\sigma_D^2. \tag{4.53}$$

This model can be approximated by the discrete Gaussian mixture model (see Figure 4.3)

$$p(D_R; \sigma_D^2) \approx \sum_{j=1}^{J} p(D_R|\sigma_{D_j}^2)w_{D_j}, \tag{4.54}$$

where

$$w_{D_j} = p(\sigma_D^2 = \sigma_{D_j}^2), \tag{4.55}$$

is now a discrete approximation of the distribution of the variance σ_D^2. Given $p(\sigma_D^2 = \sigma_{D_j}^2)$, we can compute the parameters w_{D_j} and r_{D_j} with Equations 4.55 and 4.49 respectively and thus determine the noise prior in Equation 4.7.
In real world situations the variance σ_D^2 is not available and has to be estimated as $\hat{\sigma}_D^2$. It is possible therefore to approximate $p(\hat{\sigma}_D^2 = \hat{\sigma}_{D_j}^2)$ by constructing a histogram of $\hat{\sigma}_D^2$. Each bin of the histogram would correspond to one mixture, with a weight equal to its relative frequency and a variance ratio computable from the bin centers using Equation 4.49. One disadvantage when using this method is that noise variance estimation

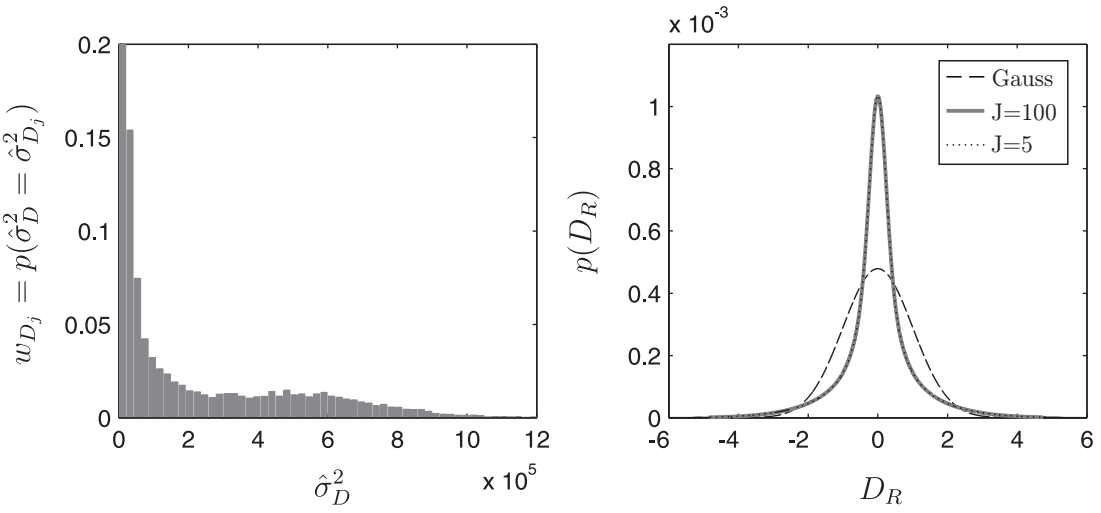

Figure 4.4: **Left:** Empirically obtained distribution of the estimated variance of the real Fourier component of noise $\hat{\sigma}_D^2$ (shaded). **Right:** Gaussian prior of the real component of noise D_R (dashed). $J = 100$ mixtures GMM noise prior of D_R corresponding to the histogram on the left (solid gray). $J = 5$ mixtures GMM noise prior trained with Montecarlo simulation (dotted black).

distributions do normally have very long tails. A high number of mixtures is therefore needed to model $p(\hat{\sigma}_D^2 = \hat{\sigma}_{D_j}^2)$ properly which implies a high computational cost.
This problem can be overcome by dividing the training into two steps.

1. A large set of samples of $\hat{\sigma}_D^2$ is obtained by using a noise variance estimator (see Section 2.2) on speech enhancement tests. A high resolution histogram (i.e. $J = 100$ bins/mixtures) is computed from these samples (Figure 4.4, left). The parameters w_{D_j}, r_{D_j} of the GMM noise prior $p(D_R)$ corresponding to that histogram are computed.

2. Draw N random samples[5] from $p(D_R)$ and train a much smaller ($J = 5$) GMM prior using the EM algorithm. The resulting prior provides a 20 fold reduction of resource consumption and is almost indistinguishable from the original (see Figure 4.4).

Provided that the supergaussian characteristics of the prior depend mainly on the method used for the estimation of variance and not on the characteristics of the noise, the method presented in this section allows us to model a specific prior for a given noise estimation technique by averaging through various test conditions. As it will be shown in the experimental setup, this was used to adapt the RiMM-MMSE estimators to the IMCRA noise variance estimator.

[5]This implies sampling from a discrete distribution of mixtures and subsequently from each Gaussian mixture.

5 Fourier Domain Uncertainty Models

The context of speech enhancement for additive noise was introduced in Chapter 2 and extended in Chapter 4. Given this context, the term 'Fourier domain uncertainty model' will be used here to denote a probabilistic description of the clean signal used to model the missing information after speech enhancement. This probabilistic description will be in form of a posterior distribution $p(X_{kl}|Y_{kl})$ of each Fourier coefficient of the clean signal X_{kl} given the observed information Y_{kl} (see Figure 5.1). This is however a notation simplification. The available information for the estimation of X_{kl} is not limited to Y_{kl}. Other noisy Fourier coefficients $Y_{k'l'}$ can also be indirectly used to determine the posterior distribution of X_{kl} (i.e. SNR estimation seen in Section 2.2).
The STFT domain has some very desirable properties. It is a linear transformation of the time domain, from which the original time domain signal can be completely recovered by using the inverse short-time Fourier transform (ISTFT). It is also a domain, in which phenomena like additive or convolutive interactions between speech and noise have simple expressions and the parameters ruling the corresponding statistical models are easy to compute. Due to this and other reasons, STFT is the domain most often chosen for the implementation of speech enhancement and source separation algorithms.

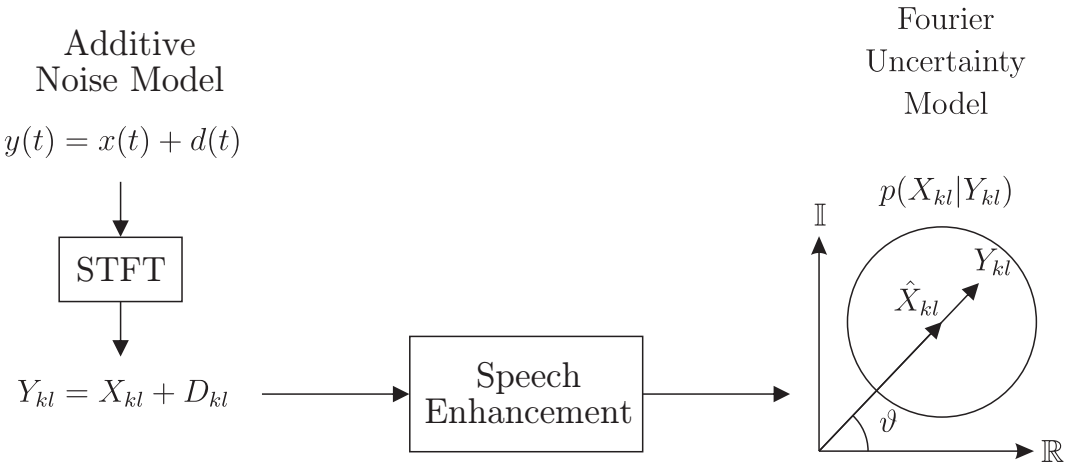

Figure 5.1: Uncertain description of the clean Fourier coefficient X_{kl} after speech enhancement using a circularly symmetric distribution of mean equal to the estimated clean Fourier coefficient $\hat{X}_{kl}$.

5.1 A Generalized Model of Uncertainty in the Fourier Domain

Uncertainty models in the STFT domain for uncertainty propagation were initially employed by Kolossa [Kolossa 2005] in the context of source separation by means of independent component analysis (ICA). For this purpose, the uncertainty of information was used to compensate for the loss of information incurred by time-frequency post-masking of ICA results. Each time frequency element of the STFT spectrum was considered as a Gaussian distributed variable, with mean equal to the amplitude of the estimated clean STFT element $|\hat{X}|$ and an associated variance λ. This uncertain description was then propagated into the cepstral domain using the log-Normal approximation and pseudo-Montecarlo methods [Kolossa 2008]. This same model was later employed to propagate the uncertainty generated after single channel speech enhancement, using the estimated noise variance [Kolossa 2006]. Gaussian uncertainty models were also employed by Srinivasan for the same purpose employing a neural network [Srinivasan 2006] and regression tress [Srinivasan 2007] for uncertainty propagation. A similar approach to that of [Kolossa 2006] was also used by Stouten to propagate noise variance statistics, rather than a posterior of the clean Fourier coefficient, in [Stouten 2006b].

5.1.1 The Complex Gaussian Model of Uncertainty

In this work, the Fourier domain uncertainty models are extended to the circularly symmetric complex Gaussian distribution (see Section 2.3.1). Under this model the k^{th} frequency bin Fourier coefficient of the l^{th} frame of the clean signal is described as a complex Gaussian distribution

$$p(X_{kl}|Y_{kl}) = \frac{1}{\pi\lambda_{kl}} \exp\left(-\frac{|X_{kl} - \hat{X}_{kl}|^2}{\lambda_{kl}}\right). \tag{5.1}$$

of mean equal to the estimated clean Fourier coefficient

$$\hat{X}_{kl} = G_{kl} \cdot Y_{kl}, \tag{5.2}$$

and variance λ_{kl}
As seen in Section 2.1 this model has some very desirable properties.

- The distribution of its amplitude is known to follow a Rice distribution (see Section 2.3.2).
- There exists a closed form formula for the n^{th} order raw moment of its amplitude (similar to Equation 2.120), from which higher oder uncertainty information can be computed.
- As it will be seen in Section 5.3, it arises naturaly as a model for uncertainty when using Wiener and MMSE-STSA estimators.

- As it will be seen in Chapter 6, it is an uncertainty model of amplitude and phase and thus it can be propagated back into the time domain, resulting in a Gaussian distributed uncertain time domain signal.
- It is simple model in the sense that it arises from the assumption of Gaussianity and statistical independence of the real and imaginary components of the signal and thus contains no higher order information (i.e. higher order cumulants are zero).

5.1.2 The Non-Central Chi and Non-Central Chi-Square Models of Uncertainty

The complex Gaussian uncertainty model can also be related to the Gaussian uncertainty model used in [Kolossa 2005, Kolossa 2006, Srinivasan 2006, Srinivasan 2007, Kolossa 2008]. Both models originate, in fact, from a more general model. Let X_1 denote a complex Gaussian distributed uncertain Fourier coefficient. This can be expressed in terms of its real and imaginary components as

$$X_1 = X_R + iX_I, \tag{5.3}$$

where X_R and X_I are Gaussian distributed with the following relation between the complex mean and the mean of its components

$$\hat{X} = \hat{X}_R + i\hat{X}_I, \tag{5.4}$$

and variance $\lambda/2$. The amplitude of this uncertain Fourier coefficient can be expressed as

$$A_1 = |X_1| = \sqrt{X_R^2 + X_I^2}, \tag{5.5}$$

Let us now consider a real valued, Gaussian distributed uncertain Fourier coefficient X_2, with mean $|\hat{X}|$ and variance $\lambda/2$. The amplitude of this random variable can be expressed as

$$A_2 = |X_2| = \sqrt{X_2^2}. \tag{5.6}$$

If both uncertain amplitudes in Equations 5.5 and 5.6 are now divided by $\sqrt{\lambda/2}$, by the definition of non-central Chi distribution [Johnson 1970b, Chap. 28]

$$\frac{A_1}{\sqrt{\lambda/2}} = \sqrt{\frac{X_R^2}{\lambda/2} + \frac{X_I^2}{\lambda/2}}, \tag{5.7}$$

follows a non-central Chi-distribution of two degrees of freedom with non-centrality parameter

$$l = \sqrt{\frac{\hat{X}_R^2}{\lambda/2} + \frac{\hat{X}_I^2}{\lambda/2}} = \sqrt{\frac{|\hat{X}|^2}{\lambda/2}}, \tag{5.8}$$

whereas

$$\frac{A_2}{\sqrt{\lambda/2}} = \sqrt{\frac{X_2^2}{\lambda/2}}, \tag{5.9}$$

follows a non-central Chi-distribution with one degree of freedom and same non-centrality parameter.
This result implies that both Gaussian and complex Gaussian uncertainty models can be treated as particular cases of a more general model, where the amplitude of the normalized Fourier coefficient follows a non-central Chi distribution and its power a non-central Chi square distribution [Johnson 1970b, Chap. 28]. This model shares some advantages with the complex Gaussian model, like the existence of closed-form solutions for their n^{th} order raw moment[1]. It also allows to set an arbitrary number of Gaussian uncertainty sources (degrees of freedom) from which the uncertainty is generated. This work will be centered however in the particular case of the complex Gaussian model due to the additional properties commented on the beginning of the chapter.

5.2 Uncertainty Estimation based on Empirical Methods

When performing speech enhancement under real world conditions, there is always a certain amount of information that is missing (i.e. the interfering signal) and thus completely recovering the original signal is an impossible task. This missing information is the source of randomness, which justifies the use of a probabilistic description of the spectrum of the signal given our observable data. Unfortunately there are multiple sources of uncertainty in real world situations and there is no unique way to determine the form or statistical parameters of the resulting randomness.
One possible method to determine this uncertainty was proposed in the context of time-frequency post-masking of ICA results in [Kolossa 2005]. This method combined the hypothesis over the two typical errors in STFT estimation, speech attenuation and residual noise, using speech probability. For periods of speech absence (H_0) it was assumed that the uncertainty variance would be proportional to the estimated clean speech amplitude (considered as residual noise)

$$\tilde{\lambda}_{kl}^{H_0} = |\hat{X}_{kl}|, \tag{5.10}$$

whereas for periods of speech presence (H_1), it was assumed that the uncertainty variance would be the result of speech attenuation occurring during speech enhancement[2]

$$\tilde{\lambda}_{kl}^{H_1} = |Y_{kl}| - |\hat{X}_{kl}|. \tag{5.11}$$

[1] Note that for the purpose of uncertainty propagation the fact that the random variables are divided by $\sqrt{\lambda/2}$ is trivially solved since it is a linear operation.

[2] In the context of ICA post-masking, gain functions are bounded between zero and one and thus they can only result in no change or attenuation of the Fourier coefficient. For details refer to [Kolossa 2005, Kolossa 2006].

Combining both hypotheses using their respective probabilities resulted in the following formula

$$\tilde{\lambda}_{kl} = p(H_0|Y)|\hat{X}_{kl}| + p(H_1|Y)(|Y_{kl}| - |\hat{X}_{kl}|). \tag{5.12}$$

Such measures depend, however, on a correct estimation of the speech probabilities and proved to be ineffective in single channel speech enhancement. Based on this idea, measures of uncertainty for single channel speech enhancement methods employing self-reference measures were developed in this work. In the absence of further information, such methods estimate the variance of the uncertain Fourier coefficient λ as proportional to the difference between the Fourier coefficients of two steps of the same speech enhancement procedure. The simplest uncertainty variance estimator using this principle is based on the difference between the input and output STFT of the clean speech estimators seen in Chapter 2

$$\tilde{\lambda}_{kl} = |Y_{kl} - \hat{X}_{kl}|^2 = (1 - G_{kl})^2|Y|^2, \tag{5.13}$$

This estimator and its variations result however in overestimations of the uncertainty, which limits their efficiency for high SNRs, particularly under the observation uncertainty methods currently used to incorporate uncertainty to the speech recognition process[3]. Furthermore, since they are ad hoc. methods, based on empirical observations, they generalize poorly to the different testing environments.
As it will be shown in the experimental setup, such methods do however offer a reasonable trade-off between simplicity and estimation accuracy when dealing with real world speech enhancement systems like the ETSI ES202 050 standard. The particular implementation used with the ETSI standard, described in the next section, employs the power subtraction estimator (see Section 2.3.3) to yield an estimation of uncertainty.

5.2.1 Estimation of Uncertainty for the ETSI Advanced Front-End

As seen in Section 3.4 and displayed in Figure 3.6 the ETSI advanced front-end combines multiple speech enhancement techniques for an optimal performance.
Since the standard shows excellent suppression of residual noises at the noise reduction block, the main source of uncertainty is assumed to be the waveform processing and pre-emphasis steps. Such steps depend on obtaining a high SNR after the noise reduction step to function properly, which might not always be the case.
The uncertainty variance was then computed as a function of the amount of change between the STFTs of the signals before and after the waveform processing and pre-emphasis steps. As seen in Figure 5.2 these correspond to $\hat{X}^N_{kl}$ and $\hat{X}_{kl}$ respectively. Given the output Fourier coefficients of these consecutive steps, we can express the change between them as

$$\hat{X}_{kl} = \hat{X}^N_{kl} + \delta_{kl}. \tag{5.14}$$

If we assume the conventional statistical model for additive noise in the STFT domain (see Chapter 2), where δ_{kl} and $\hat{X}^N_{kl}$ are zero mean complex Gaussian distributed, and

[3]This aspect will be discussed in the conclusions in Section 8.1.2

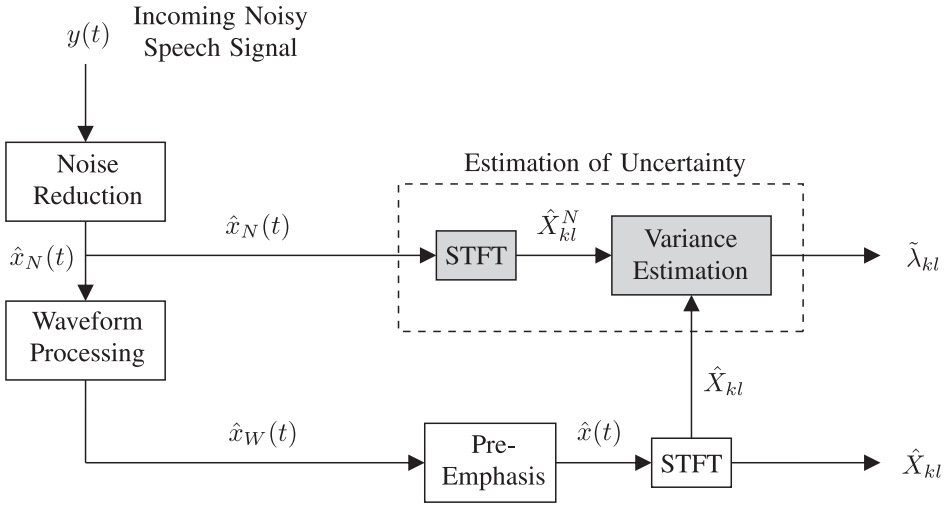

Figure 5.2: Extension of the ETSI ES 202 050 standard for the estimation of uncertainty. Only the speech enhancement blocks and the needed modifications are displayed.

we use the fact that $\hat{X}_{kl}^N$ and $\hat{X}_{kl}$ are observable processes, we can estimate the variance of the hidden process $\lambda_{\delta_{kl}}$ using a conventional power subtraction estimator (see Section 2.3.3) as

$$\hat{\lambda}_{\delta_{kl}}^{\mathrm{ML}} = \arg\max_{\lambda_{\delta_{kl}}} \left\{ p(\hat{X}_{kl}; \lambda_{\delta_{kl}}) \right\} = \arg\max_{\lambda_{\delta_{kl}}} \left\{ \frac{1}{\pi(\hat{\lambda}_{X_{kl}^N} + \lambda_{\delta_{kl}})} \exp\left(-\frac{|\hat{X}_{kl}|^2}{\hat{\lambda}_{X_{kl}^N} + \lambda_{\delta_{kl}}} \right) \right\} \tag{5.15}$$

where the variance $\lambda_{X_{kl}^N}$ can be obtained as maximum likelihood estimate of the variance of the prior $p(\hat{X}_{kl}^N; \lambda_{X_{kl}^N})$ from the observed Fourier coefficient $\hat{X}_{kl}^N$ as

$$\hat{\lambda}_{X_{kl}^N}^{\mathrm{ML}} = \arg\max_{\lambda_{X_{kl}^N}} \left\{ \frac{1}{\pi \lambda_{X_{kl}^N}} \exp\left(-\frac{|\hat{X}_{kl}^N|^2}{\lambda_{X_{kl}^N}} \right) \right\} = |\hat{X}_{kl}^N|^2. \tag{5.16}$$

By replacing Equation 5.16 in Equation 5.15, we obtain

$$\hat{\lambda}_{\delta_{kl}}^{\mathrm{ML}} = \arg\max_{\lambda_{\delta_{kl}}} \left\{ \frac{1}{\pi(|\hat{X}_{kl}^N|^2 + \lambda_{\delta_{kl}})} \exp\left(-\frac{|\hat{X}_{kl}|^2}{|\hat{X}_{kl}^N|^2 + \lambda_{\delta_{kl}}} \right) \right\} = |\hat{X}_{kl}|^2 - |\hat{X}_{kl}^N|^2 \tag{5.17}$$

As in the conventional power subtraction algorithm, the resulting estimate of $\lambda_{\delta_{kl}}$ has to be constrained to be greater than zero leading to the following approximation of the estimation uncertainty

$$\left.\begin{array}{l} p(X) = \frac{1}{\pi\lambda_X}\exp\left(-\frac{|X|^2}{\lambda_X}\right) \\ p(D) = \frac{1}{\pi\lambda_D}\exp\left(-\frac{|D|^2}{\lambda_D}\right) \\ Y = X + D \end{array}\right\} \Rightarrow$$

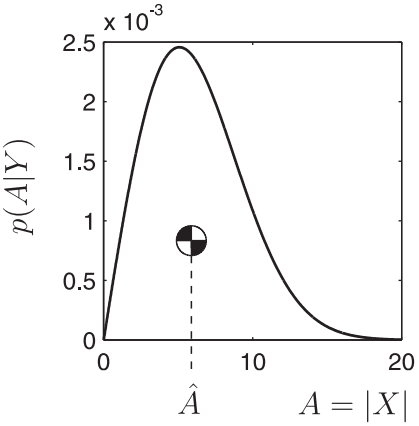

Figure 5.3: Posterior distribution of the short-time spectral amplitude for the conventional Gaussian model under the assumption of additive noise.

$$\tilde{\lambda}_{kl} \approx \max\left\{|\hat{X}_{kl}|^2 - |\hat{X}_{kl}^N|^2, 0\right\}. \tag{5.18}$$

For conventional implementations of the MMSE estimators, like the ones introduced in Chapters 2 and 4, a method for uncertainty estimation is presented in the next section.

5.3 Posterior Distribution as a Measure of Uncertainty

5.3.1 MMSE-STSA Posterior Distribution

As seen in the introduction to speech enhancement (see Chapter 2), MMSE methods provide an estimation of the clean signal by minimizing an expected measure of the Euclidean estimation error. The solution to this minimization problem was also demonstrated to correspond to the expectation, or center of probabilistic density, of the posterior distribution. For example, the MMSE estimation of the short-time spectral amplitude of a Fourier coefficient of the clean signal (MMSE-STSA estimator) corresponds to

$$\hat{A} = \arg\min_{\hat{A}}\left\{E\left\{(A-\hat{A})^2\right\}\right\} = E\{A|Y\}. \tag{5.19}$$

where, as shown in Figure 5.3, $\hat{A}$ corresponds to the center of probabilistic density or centroid of the posterior distribution derived in Section 4.3.3

$$p(A|Y) = \frac{p(Y|A)p(A)}{p(Y)} = \frac{2A}{\lambda}\exp\left(-\frac{\left|\frac{\lambda_X}{\lambda_X+\lambda_D}Y\right|^2 + A^2}{\lambda}\right) I_0\left(\frac{2\left|\frac{\lambda_X}{\lambda_X+\lambda_D}Y\right|A}{\lambda}\right), \tag{5.20}$$

where

$$\lambda = \frac{\lambda_X \lambda_D}{\lambda_X + \lambda_D}, \tag{5.21}$$

was defined in Equation 4.28. The posterior distribution $p(A|Y)$ can then be considered as a probabilistic description of the value of the clean signal Fourier coefficient amplitude A given the observable data Y and the assumed statistical model. Furthermore, this posterior can be directly related to the mean square error (MSE)[4] of estimation by noting that

$$E\left\{(A-\hat{A})^2\right\} = \int\left(\int (A - E\{A|Y\})^2 p(A|Y)dX\right)p(Y)dY = \int \mathrm{Var}\{A|Y\}p(Y)dY. \tag{5.22}$$

However, the computation of the estimation error requires solving the integral above. For the case of the MMSE-STSA and MMSE-LSA estimators, no exact closed-form solution is known.

5.3.2 Wiener Filter Posterior Distribution

Let us however consider the MMSE estimator under the same model but in the complex domain. The Wiener filter derived in Section 2.1.4 corresponds to the independent MMSE estimation of the real and imaginary components of the clean signal. The minimization problem for the real component, for example, leads to

$$\hat{X}_R = \arg\min_{\hat{X}_R}\left\{E\left\{(X_R - \hat{X}_R)^2\right\}\right\} = E\{X_R|Y_R\}. \tag{5.23}$$

The corresponding posterior distribution can be computed as

$$p(X_R|Y_R) = \frac{p(Y_R|X_R)p(X_R)}{p(Y_R)}, \tag{5.24}$$

where $p(X_R)$, $p(Y_R|X_R)$, $p(Y_R)$, are defined in Equations 2.96, 2.97 and 2.98 respectively. The solution to this posterior is straightforward since

$$p(X_R|Y_R) = \frac{1}{\sqrt{\pi\frac{\lambda_X\lambda_D}{\lambda_X+\lambda_D}}}\exp\left(-\frac{X_R^2}{\lambda_X} - \frac{(Y_R - X_R)^2}{\lambda_D} + \frac{Y_R^2}{\lambda_D+\lambda_X}\right), \tag{5.25}$$

where the exponent can be reorganized to yield the exponent of a conventional Gaussian distribution

[4]Note that here MSE refers to the Bayesian mean square error since it is only used in the context of MMSE estimation.

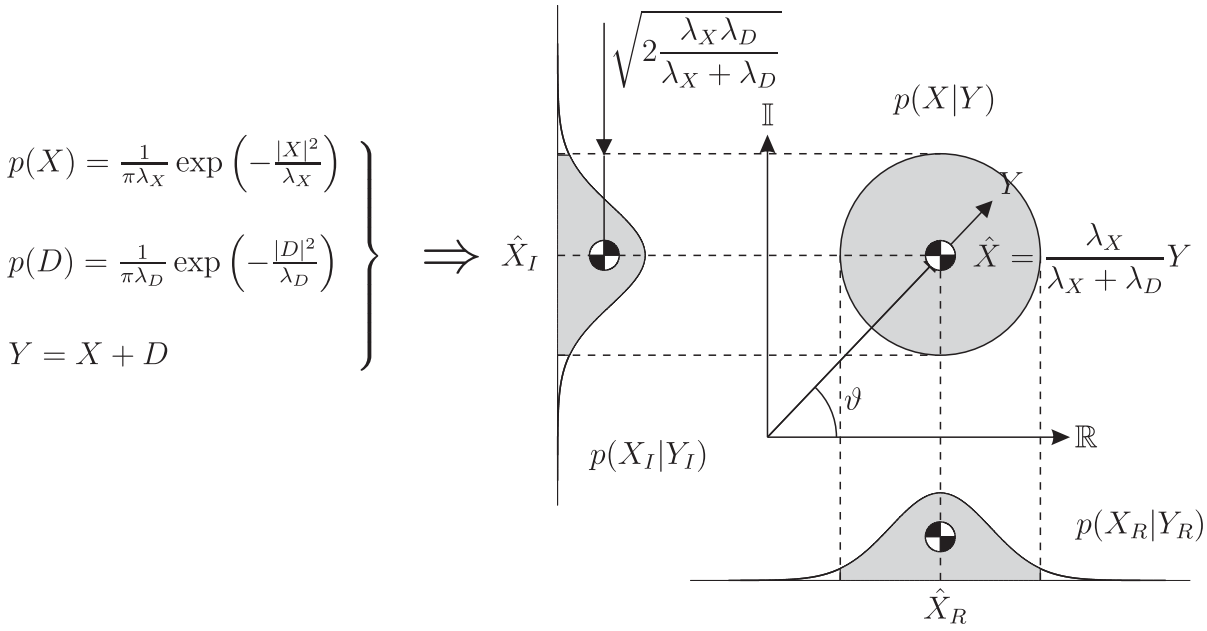

Figure 5.4: Posterior distribution of the short-time Fourier coefficient for the conventional Gaussian model under the assumption of additive noise.

$$
\begin{aligned}
-\frac{X_R^2}{\lambda_X} - \frac{(Y_R - X_R)^2}{\lambda_D} + \frac{Y_R^2}{\lambda_D + \lambda_X} = \\
-\frac{\lambda_X + \lambda_D}{\lambda_X \lambda_D} \left((\lambda_X + \lambda_D) \frac{X_R^2}{\lambda_X + \lambda_D} + \frac{\lambda_X^2 Y_R^2}{(\lambda_X + \lambda_D)^2} - 2\frac{\lambda_X Y_R X_R}{\lambda_D + \lambda_X} \right) = \\
-\frac{\lambda_X + \lambda_D}{\lambda_X \lambda_D} \left(X_R - \frac{\lambda_X}{\lambda_X + \lambda_D} Y_R \right)^2, \qquad (5.26)
\end{aligned}
$$

leading to

$$
p(X_R|Y_R) = \frac{1}{\sqrt{\pi\lambda}} \exp\left(-\frac{\left(X_R - \frac{\lambda_X}{\lambda_X+\lambda_D} Y_R \right)^2}{\lambda} \right). \qquad (5.27)
$$

Following this same procedure, the equivalent posterior for the imaginary part can be computed as

$$
p(X_I|Y_I) = \frac{1}{\sqrt{\pi\lambda}} \exp\left(-\frac{\left(X_I - \frac{\lambda_X}{\lambda_X+\lambda_D} Y_I \right)^2}{\lambda} \right). \qquad (5.28)
$$

Since under the assumed model real and imaginary components of the Fourier coefficients Y and X are considered statistically independent, their corresponding posteriors are

also statistically independent and the posterior of the clean Fourier coefficient can be computed as (see Figure 5.4)

$$p(X|Y) = p(X_R, X_I|Y_R, Y_I) = p(X_R|Y_R)p(X_I|Y_I) = \frac{1}{\pi\lambda}\exp\left(-\frac{\left|X - \frac{\lambda_X}{\lambda_X+\lambda_D}Y\right|^2}{\lambda}\right) = \frac{1}{\pi\lambda}\exp\left(-\frac{\left|X - X^W\right|^2}{\lambda}\right), \tag{5.29}$$

where the definition of circularly symmetric complex Gaussian distribution introduced in Section 2.3.1 was used and X^W is the solution to the Wiener filter

$$\hat{X}^W = \frac{\lambda_X}{\lambda_X + \lambda_D}Y. \tag{5.30}$$

As it was done in the Ephraim-Malah filter case, the MSE in the complex domain can be computed as

$$E\left\{|X - \hat{X}^W|^2\right\} = \int\left(\int |X - E\{X|Y\}|^2 p(X|Y)dX\right)p(Y)dY = \int \mathrm{Var}\{X|Y\}p(Y)dY = \int \lambda p(Y)dY = \lambda, \tag{5.31}$$

where in this case the integral has a simple closed form solution, since the variance of the posterior does not depend on the Fourier coefficient of the noisy signal Y. The variance of the Wiener filter is in fact equal to the Bayesian MSE.

5.3.3 An Approximated Posterior When Combining MMSE Amplitude Estimators with the Optimal Phase Estimator

Given the result in Equation 5.31, it would also be desirable to find a similar closed form expression for the Fourier coefficient estimation error obtained when combining the optimal phase estimator (see Equation 2.40) with the MMSE-STSA amplitude estimator (see Equation 2.37)

$$\hat{X}^{\mathrm{STSA}} = \hat{A}^{\mathrm{STSA}}\frac{Y}{|Y|} = \Gamma(1.5)\frac{\sqrt{\nu}}{\gamma}\exp\left(-\frac{\nu}{2}\right)\left[(1+\nu)I_0\left(\frac{\nu}{2}\right) + \nu I_1\left(\frac{\nu}{2}\right)\right]Y, \tag{5.32}$$

and the MMSE-LSA amplitude estimator (see Equation 2.46)

$$\hat{X}^{\mathrm{LSA}} = \hat{A}^{\mathrm{LSA}}\frac{Y}{|Y|} = \frac{\xi}{1+\xi}\exp\left(\frac{1}{2}\int_\nu^\infty \frac{e^{-t}}{t}dt\right)Y. \tag{5.33}$$

In order to do this, we observe the dependency of both estimators on the parameter

$$\nu = \frac{\xi}{1+\xi}\gamma = \frac{\left(\frac{\lambda_X}{\lambda_X+\lambda_D}|Y|\right)^2}{\frac{\lambda_X\lambda_D}{\lambda_X+\lambda_D}} = \frac{|X^W|^2}{\lambda}. \tag{5.34}$$

This parameter corresponds to the ratio r of the MMSE-STSA Rice distributed posterior in Equation 5.20 (see the definition in Section 2.3.2, Equation 2.77).
As indicated in [Ephraim 1984, Eqs. 11, 12], the confluent hypergeometric function in Equation 2.116 has the following asymptotic behavior for $\nu >> 1$

$$\Phi\left(-0.5; 1; -\nu\right) \approx \frac{\sqrt{\nu}}{\Gamma(1.5)}. \tag{5.35}$$

By replacing this in Equation 2.116, it is clear that the MMSE-STSA amplitude estimator converges to the Wiener filter solution for $\nu >> 1$

$$\hat{X}^{\text{STSA}} \approx \frac{\xi}{1+\xi} Y = \hat{X}^{W}. \tag{5.36}$$

In the case of the MMSE-LSA estimator

$$\hat{X}^{\text{MMSE-LSA}} = \frac{\xi}{1+\xi} \exp\left(\frac{1}{2}\int_{\nu}^{\infty} \frac{e^{-t}}{t} dt\right) Y, \tag{5.37}$$

Martin's approximation for the exponential integral function (see Equation 2.128) can be used to yield the following for $\nu > 1$

$$\int_{\nu}^{\infty} \frac{e^{-t}}{t} dt \approx 10^{-0.52\nu - 0.26}, \tag{5.38}$$

which tends to zero for $\nu >> 1$, thus, leading to the same asymptotic behavior as the MMSE-STSA estimator

$$\hat{X}^{\text{LSA}} \approx \frac{\xi}{1+\xi} Y = \hat{X}^{W}. \tag{5.39}$$

In addition to this convergence, the estimator for the phase of the clean Fourier coefficient α, used in Equations 5.32 and 5.33 coincides with that of the Wiener filter. Furthermore, the amplitude distribution of the complex Gaussian Wiener posterior in Equation 5.29 coincides with the posterior of the MMSE-STSA estimator Equation 5.20. For a high value of ν it can be therefore assumed that the distribution of the clean Fourier coefficient $p(X|Y)$ for the MMSE-STSA Fourier coefficient estimation is good approximated by the posterior distribution of the Wiener filter in Equation 5.29 but with mean equal to X^{STSA}, rather than X^{W}, leading to the following approximation

$$p(X|Y) \approx \frac{1}{\pi\lambda} \exp\left(-\frac{\left|X - \hat{X}^{\text{STSA}}\right|^2}{\lambda}\right). \tag{5.40}$$

Regarding the estimation of the Fourier coefficient using the MMSE-LSA estimator, the form of the posterior distribution of the log-amplitude $p(\log A|Y)$ is not known. However, this probabilistic description of the log-amplitude is propagated back into the the amplitude domain by using the exponential (see Equation 2.44). Consequently the same approximation as in the case of the MMSE-STSA is used for the Fourier domain posterior when the MMSE-LSA estimator is used

$$p(X|Y) \approx \frac{1}{\pi\lambda} \exp\left(-\frac{\left| X - \hat{X}^{\mathrm{LSA}} \right|^2}{\lambda} \right). \tag{5.41}$$

For the cases in which the value of ν is not very high, Wiener and Ephraim-Malah filter solutions differ. If we assume, however, that this difference remains small compared to the variance of the Wiener filter λ, we can still use this approximation in such cases.

5.3.4 Supergaussian Measures of Uncertainty

In a similar way as it was done for the Gaussian model, a posterior distribution of each Fourier coefficient given the observable information can be derived for the Gaussian mixture model presented in Chapter 4. Nevertheless this can not be interpreted as the posterior of a Wiener estimator, since the prior distributions are not Gaussian distributed.
Given the complex Gaussian mixture model likelihood $p(Y|X)$ in Equation 4.6 and prior $p(X)$ in Equation 4.10, a probabilistic description of the clean signal can be obtained through the Bayes theorem. This can be attained by using Equations 4.31 and 4.32 to express the Gaussian mixture models as

$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} p(i)p(j)p(Y|X,j)p(X|i)}{\sum_{i=1}^{I}\sum_{j=1}^{J} p(i)p(j)p(Y|i,j)}. \tag{5.42}$$

Reorganizing the exponents, this yields a mixture of the complex Gaussian posteriors obtained in the Gaussian case for the Wiener filter (see Equation 5.29),

$$p(X|Y) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{p(i)p(j)p(Y|i,j)}{\sum_{i=1}^{I}\sum_{j=1}^{J} p(i)p(j)p(Y|i,j)} p(X|Y,i,j), \tag{5.43}$$

where the weights are the same as those of the Rice mixture posterior in Equation 4.29 and the $(i,j)^{th}$ mixture

$$p(X|Y,i,j) = \frac{p(Y|X,j)p(X|i)}{p(Y|i,j)}, \tag{5.44}$$

corresponds to the complex Gaussian distributed posterior of a Wiener filter (see Section 5.3.2) evaluated for the parameters Y, λ_{X_i}, λ_{D_j}, thus giving

$$p(X|Y) = \sum_{i=1}^{I}\sum_{j=1}^{J} \Omega_{ij} p(X|Y,i,j) = \sum_{i=1}^{I}\sum_{j=1}^{J} \Omega_{ij} \frac{1}{\pi\lambda_{ij}} \exp\left(\frac{\left| X - \frac{\lambda_{X_i}}{\lambda_{X_i}+\lambda_{D_j}} Y \right|^2}{\lambda_{ij}} \right), \tag{5.45}$$

where Ω_{ij} was defined in Equation 4.30 and λ_{ij} is defined in Equation 4.28.

6 Propagation of Uncertainty through the Feature Extraction Process

6.1 Piecewise Uncertainty Propagation

There are many possible methods to compute the propagation of a random variable through a non-linear transformation, from interpolation to variational methods. However, there are two characteristics of the particular problem of uncertainty propagation for robust automatic speech recognition, that limit the techniques usable

1. ASR systems are expected to work in real time or with reasonable offline execution times. Computational resources are also scarce in some particular cases (see ETSI standard in Section 3.4).

2. Feature extractions combine linear and non-linear operations performed jointly on multiple features of the same frame or combining features from different time frames. Such transformations can therefore become very complex.

Regarding the first characteristic, the fastest solution to a propagation problem is finding a closed-form solution for the PDF or the needed cumulants of the transformed random variable. This implies solving changes of variable, derivatives or integrals which due to the complexity of the feature extraction transformations are rarely obtainable. Regarding the second characteristic, pseudo-Montecarlo methods allow the estimation of some cumulants of the transformed variable by transforming sets of characteristic samples of the distribution and are thus not dependent on the complexity of the transformations involved. Such methods are however too slow for the propagation of a probabilistic model of the STFT of the clean signal, like the one presented in Chapter 5[1].
Given this context, finding a general solution for the uncertainty propagation problem seems impossible. Relaxing the problem to propagation of first and second order information might simplify the task but it remains unfeasible to find a general solution.
There are however well known solutions for some non-linear transformations of given PDFs. It is also often the case that different feature extractions share the same or similar transformations. Taking this into account, the approach proposed here is a piecewise one, in which the feature extraction will be divided into different steps and the optimal method to propagate the statistical information through each step will be selected. The main disadvantage of this technique is that, even if we want to propagate only first and second order moments, we still need to consider the complete PDFs of the uncertain

[1]Note that a typical STFT frame of 20-30ms is represented by 256-512 Fourier coefficients, each of which is modeled by a single random variable.

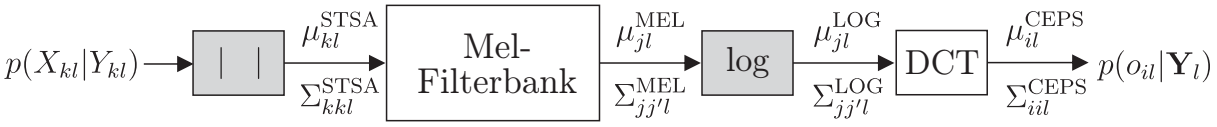

Figure 6.1: Mel-cepstral feature extraction starting in the STFT domain extended to propagate second order information into the cepstral domain. Blocks corresponding to non-linear transformations are displayed in gray. Note that only diagonal covariances are considered in the initial (STSA) and final transformations (CEPS).

features in the intermediate steps. On the positive side, we can optimize each process individually and share transformations among different feature extractions. The next sections will detail the solutions for the different transformations involved in the Mel-cepstral and RASTA-PLP feature extractions, as well as other transformations also used in feature extraction for ASR. All transformations were introduced in Chapter 3 and are detailed in the same order.

6.2 Uncertainty Propagation for the Mel-cepstral Features

6.2.1 Propagation through the STSA

If we consider the complex Gaussian Fourier uncertainty model introduced in Chapter 5, we have that each Fourier coefficient of the clean signal X_{kl} is distributed as the following complex Gaussian model

$$p(X_{kl}|Y_{kl}) = \frac{1}{\pi\lambda_{kl}} \exp\left(-\frac{|X_{kl} - \hat{X}_{kl}|^2}{\lambda_{kl}}\right), \tag{6.1}$$

where the mean is equal to the estimated clean Fourier coefficient

$$\hat{X}_{kl} = G_{kl} \cdot Y_{kl}, \tag{6.2}$$

and the variance λ_{kl} has been determined either with empirical methods or closed-form solutions (see Sections 5.2 and 5.3). As seen in Sections 2.3.1 and 2.3.2, a complex Gaussian distribution has a Rice distributed amplitude A_{kl}, in this case corresponding to

$$p(A_{kl}|Y_{kl}) = \frac{2A_{kl}}{\lambda_{kl}} \exp\left(-\frac{A_{kl}^2 + |\hat{X}_{kl}|^2}{\lambda_{kl}}\right) I_0\left(\frac{2A_{kl}|\hat{X}_{kl}|}{\lambda_{kl}}\right), \tag{6.3}$$

for which the n^{th} order moment has the following closed form solution (see derivations in Section 2.3.7 and Equation 2.109)

$$E\{|X_{kl}|^n\} = \Gamma\left(\frac{n}{2}+1\right)(\lambda_{kl})^{\frac{n}{2}}\,\Phi\left(-\frac{n}{2};1;-\frac{|\hat{X}_{kl}|^2}{\lambda_{kl}}\right) = \Gamma\left(\frac{n}{2}+1\right)(\lambda_{kl})^{\frac{n}{2}}\,\mathrm{L}_{\frac{n}{2}}\left(-\frac{|\hat{X}_{kl}|^2}{\lambda_{kl}}\right). \quad (6.4)$$

and the last equality is obtained using the equivalence of the confluent hypergeometric function $\Phi()$ and the Laguerre polynomials[2] $\mathrm{L}^0_{\frac{n}{2}}()$ as given in [Gradshteyn 2007, Eq. 8.972.1].
Given a general closed-form solution of the n^{th} order moment, the mean of the uncertain STSA can be computed as

$$\mu_{kl}^{\mathrm{STSA}} = E\{|X_{kl}|\} = \Gamma(1.5)\sqrt{\lambda_{kl}}\mathrm{L}_{\frac{1}{2}}\left(-\frac{|\hat{X}_{kl}|^2}{\lambda_{kl}}\right) \quad (6.5)$$

where $\mathrm{L}_{\frac{1}{2}}()$ can be expressed in terms of Bessel functions by combining [Gradshteyn 2007, Eq. 8.972.1] and [Rice 1944, Eq. 4B-9] as

$$\mathrm{L}_{\frac{1}{2}}(x) = \exp\left(\frac{x}{2}\right)\left[(1-x)\,I_0\left(-\frac{x}{2}\right) - xI_1\left(-\frac{x}{2}\right)\right]. \quad (6.6)$$

Note the logical similarity between Equation 6.5 and the MMSE-STSA estimate given by Equation 2.37, since both correspond to the mean amplitude of a complex Gaussian distribution. Note also that for a variance λ_{kl} equal to that of the Wiener filter posterior (see Section 5.3.2, Equation 5.31) both formulas coincide.
The variance of the uncertain STSA can be computed in a similar form, using the relation between second order cumulants and moments [Chrysostomos 1993, Eq. 2.4] as

$$\Sigma_{kkl}^{\mathrm{STSA}} = E\{|X_{kl}|^2\} - E\{|X_{kl}|\}^2, \quad (6.7)$$

where the second order moment can be obtained from Equation 6.4 using [Gradshteyn 2007, Eq. 8.970.4] as

$$E\{|X_{kl}|^2\} = \lambda_{kl} + |\hat{X}_{kl}|^2, \quad (6.8)$$

leading to

$$\Sigma_{kkl}^{\mathrm{STSA}} = \lambda_{kl} + |\hat{X}_{kl}|^2 - \left(\mu_{kl}^{\mathrm{STSA}}\right)^2. \quad (6.9)$$

The covariance of each STSA frame $\mathbf{A}_l$ is diagonal since each Fourier coefficient is considered statistically independent of the others.
Although first and second order information suffice to continue with the uncertainty propagation approach used here, higher order moments are also available from Equation 6.4. Particularly interesting are the third and fourth order cumulants from which

[2]Note that the upper index in the Laguerre polynomial will be omitted from now on.

skewness and kurtosis can be computed. The third order cumulant can be obtained using the relation between the third order cumulant and moments [Chrysostomos 1993, Eq. 2.4] as

$$\begin{aligned} \mathrm{K3}_{kl}^{\mathrm{STSA}} = E\left\{|X_{kl}|^3\right\} - 3E\left\{|X_{kl}|^2\right\} E\left\{|X_{kl}|\right\} - 2E\left\{|X_{kl}|\right\}^3 = \\ E\left\{|X_{kl}|^3\right\} - 3\Sigma_{kkl}^{\mathrm{STSA}}\mu_{kl}^{\mathrm{STSA}} - \left(\mu_{kl}^{\mathrm{STSA}}\right)^3, \end{aligned} \tag{6.10}$$

where the third order raw moment corresponds to

$$E\left\{|X_{kl}|^3\right\} = \Gamma(2.5)\lambda_{kl}^{\frac{3}{2}}\mathrm{L}_{\frac{3}{2}}\left(-\frac{|\hat{X}_{kl}|^2}{\lambda_{kl}}\right), \tag{6.11}$$

and the Laguerre polynomial $\mathrm{L}_{\frac{3}{2}}()$ corresponds to [Rice 1944, Ap. 4B]

$$\mathrm{L}_{\frac{3}{2}}(x) = \exp\left(\frac{x}{2}\right)\left[(3 - 6x + 2x^2)I_0\left(-\frac{x}{2}\right) + (2x^2 - 4x)I_1\left(-\frac{x}{2}\right)\right]. \tag{6.12}$$

The fourth order cumulant can be obtained similarly by applying [Gradshteyn 2007, Eq. 8.970.5] together with [Chrysostomos 1993, Eq. 2.4][3], giving

$$\mathrm{K4}_{kl}^{\mathrm{STSA}} = \lambda_{kl}^2 + 2\lambda_{kl}|\hat{X}_{kl}|^2. \tag{6.13}$$

6.2.2 Propagation through the Mel-Filterbank

The propagation through the Mel-filterbank transformation is much simpler, since it is a linear transformation (see Equation 3.17). In general, given a random vector variable $\mathbf{x}$ and a linear transformation defined by the matrix $\mathbf{T}$, the transformed mean and covariance correspond to

$$E\left\{\mathbf{T}\mathbf{x}^T\right\} = \mathbf{T}E\left\{\mathbf{x}\right\}^T, \tag{6.14}$$

and

$$\mathrm{Cov}\left\{\mathbf{T}\mathbf{x}^T\right\} = \mathbf{T}\mathrm{Cov}\left\{\mathbf{x}\right\}\mathbf{T}^T. \tag{6.15}$$

If $\mathrm{Cov}\{\mathbf{x}\}$ is diagonal, the variances of the resulting variable can be obtained as

$$\mathrm{Var}\left\{\mathbf{T}\mathbf{x}^T\right\} = \mathbf{T}'\mathrm{Var}\left\{\mathbf{x}\right\}^T, \tag{6.16}$$

with

[3]Note that this can also be obtained as the second order cumulant of a non-central Chi square distribution of two degrees of freedom scaled by $\lambda_{kl}/2$. See the generalized Chi square Fourier uncertainty model in Section 5.1.2.

$$T'_{ij} = T^2_{ij}. \tag{6.17}$$

Furthermore, if the variable $\mathbf{x}$ is multivariate Gaussian distributed, the transformed variable $\mathbf{y}$ retains this condition. If the variable is non-Gaussian but higher order cumulants are available, the cumulants of the transformed variable can be easily computed. The n^{th} order cumulant would correspond to

$$\mathrm{Cum}_n\left\{\mathbf{T}\mathbf{x}^T\right\} = \mathbf{T}'\mathrm{Cum}_n\left\{\mathbf{x}\right\}^T, \tag{6.18}$$

with

$$T'_{ij} = (T_{ij})^n. \tag{6.19}$$

In the case of the Mel-filterbank considered, according to its definition in Equation 3.17, the mean $\boldsymbol{\mu}_l^{\mathrm{Mel}}$ and covariance $\boldsymbol{\Sigma}_l^{\mathrm{Mel}}$ of the uncertain features after the Mel-filterbank transformation for each frame can be solved using Equations 6.14 and 6.15 for $\mathbf{T} = \mathbf{W}$ yielding

$$\mu_{jl}^{\mathrm{Mel}} = \sum_{k=1}^{K} W_{jk}\mu_{kl}^{\mathrm{STSA}} \tag{6.20}$$

and

$$\Sigma_{jj'l}^{\mathrm{Mel}} = \sum_{k=1}^{K}\sum_{k'=1}^{K} W_{jk}W_{j'k'}\Sigma_{kk'l}^{\mathrm{STSA}} = \sum_{k=1}^{K} W_{jk}W_{j'k}\Sigma_{kkl}^{\mathrm{STSA}} \tag{6.21}$$

where the last equality was obtained by using the fact that the STSA covariance is diagonal. It has to be taken into account that, as shown in Figure 3.3, the Mel-filters overlap. Consequently, the filterbank matrix $\mathbf{W}$ is not diagonal and thus $\boldsymbol{\Sigma}_l^{\mathrm{Mel}}$ is not diagonal as well.

6.2.3 Propagation through the Logarithm of the Mel-STSA

After the Mel-filterbank, the uncertain features of each frame follow an unknown distribution (sum of Rice random variables of different parameters[4]) and are no longer statistically independent. The calculation of a closed-form solution for the resulting PDF seems therefore unfeasible. The propagation through the logarithm domain was already solved in [Kolossa 2005] by approximating the Mel-STSA uncertain features with the log-Normal distribution. This same distribution had also been originally used in [Gales 1995] to propagate a model of corrupted speech from spectral to logarithm domain and later in [Stouten 2006b] for the propagation of noise variance statistics. However, as it will be shown in the experimental setup in Chapter 7, this assumption is sub-optimal. In the case of the complex-Gaussian Fourier uncertainty model, the empirically obtained PDF of the Mel-STSA uncertain features is closer to the Gaussian

[4]Note that since $W_{jk} > 0$ the weights can be introduced into the amplitude operator as a linear transformation on Equation 5.1.

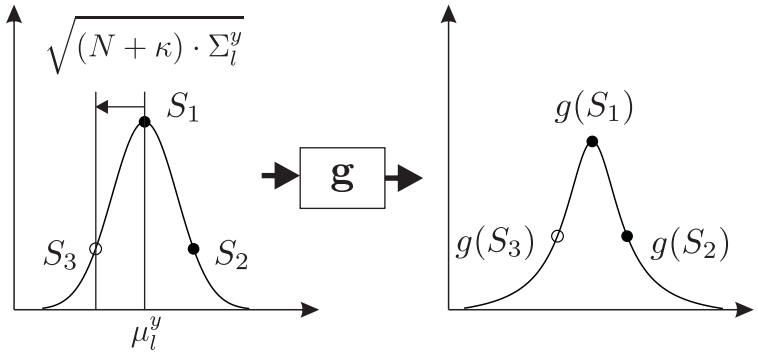

Figure 6.2: Simplified diagram of the unscented transform for the one dimensional case ($N = 1$). It is sufficient to propagate $2N + 1 = 3$ so called sigma points through the applied non-linear feature extraction $\mathbf{g}$ to compute the transformed mean and covariance.

distribution rather than the log-Normal distribution.
The fact that the distribution of the features is no longer known but it exhibits a relatively low skewness, together with the fact that, after the Mel-filterbank transformation, the number of features per frame has been reduced by almost one order of magnitude[5] favors the use of the pseudo-Montecarlo method known as unscented transform [Julier 1996]

6.2.3.1 The Unscented Transform

Montecarlo and pseudo-Montecarlo methods provide a numerical solution to the propagation problem when a closed form solution can not be found. These methods are based on replacing the continuous distribution by a set of sample points that are representative of the distribution characteristics. Statistical parameters like mean and variance can be simply computed from the transformed samples using conventional estimators. Multiple techniques are available to optimize the PDF sampling process in order to minimize the number of iterations needed and maximize accuracy. In the context of this work, however, most methods still have prohibitive execution time requirements.
The unscented transform is a pseudo-Montecarlo method that has very low computational requirements in exchange for a reasonable accuracy, particularly when some conditions about the initial and final uncertainty distribution as well as the non-linear transformation applied are met. Let

$$\mathbf{z}_l = \mathbf{g}(\mathbf{y}_l) \tag{6.22}$$

be a non-linear vector valued function. Let $\mathbf{y}_l$ be a frame containing a size N multivariate random variable of distribution $p(\mathbf{y}_l)$ with mean and covariance matrix equal to $\boldsymbol{\mu}_l^y$ and $\boldsymbol{\Sigma}_l^y$, respectively. The result of transforming $\mathbf{y}_l$ through $\mathbf{g}()$ is the multivariate random

[5]Typically the filterbank compresses each 256-512 frequency bin frame $\mathbf{X}_l$ into a compact representation of 20-25 filter outputs $\mathbf{M}_l$.

variable $\mathbf{z}_l$ of distribution $p(\mathbf{z}_l)$. The unscented transform approximates the first and second oder moments of $p(\mathbf{z}_l)$ by transforming a set of $2N+1$, so called sigma points $\mathbf{S} = \{\mathbf{S}_1 \cdots \mathbf{S}_{2N+1}\}$ which capture the characteristics of $p(\mathbf{y}_l)$(see Figure 6.2). The mean and covariance of $\mathbf{z}_l$ can be approximated by using the following weighted averages

$$E\{\mathbf{z}_l\} \approx \sum_{i=1}^{2N+1} W_i \cdot \mathbf{g}(\mathbf{S}_i) \tag{6.23}$$

$$\mathrm{Cov}\{\mathbf{z}_l\} \approx \sum_{i=2}^{2N+1} W_i \cdot (\mathbf{g}(\mathbf{S}_i) - E\{\mathbf{z}_l\}) \cdot (\mathbf{g}(\mathbf{S}_i) - E\{\mathbf{z}_l\})^T. \tag{6.24}$$

The sigma points $\mathbf{S}$ and the weights are deterministically chosen given the mean and variance of the initial probability distribution as

$$\begin{cases} W_1 = & \frac{\kappa}{(N+\kappa)} \\ \mathbf{S}_1 = & \boldsymbol{\mu}_l^y \\ \mathbf{S}_i = & \boldsymbol{\mu}_l^y + \left(\sqrt{(N+\kappa)\cdot\boldsymbol{\Sigma}_l^y}\right)_i \\ \mathbf{S}_{i+N} = & \boldsymbol{\mu}_l^y - \left(\sqrt{(N+\kappa)\cdot\boldsymbol{\Sigma}_l^y}\right)_i \\ W_i = & \frac{1}{2(N+\kappa)} \quad \text{for} \quad i \in \{2 \cdots N+1\}, \end{cases} \tag{6.25}$$

where $(\ )_i$ corresponds to the i^{th} row of the matrix square root. The additional parameter κ allows the kurtosis and higher order cumulants of $p(\mathbf{y}_l)$ to be modeled. However for multivariate distributions there are various values of kurtosis available and the option chosen was to assume a Gaussian distribution for $p(\mathbf{y}_l)$ leading to

$$\kappa = 3 - N. \tag{6.26}$$

The unscented transform is not impervious to the curse of dimensionality and suffers from accuracy problems if some conditions are not met. The assumed condition of zero odd moments for the initial distribution $p(\mathbf{y}_l)$ can be problematic in the context of uncertainty propagation. In the particular case of the Mel-STSA uncertain features, the Unscented transform does however provide a more accurate estimate than the log-Normal assumption and is therefore used, yielding

$$\mu_{jl}^{\mathrm{LOG}} \approx \sum_{i=1}^{2N+1} W_i \cdot \log(S_{ji}) \tag{6.27}$$

$$\Sigma_{jj'l}^{\mathrm{LOG}} \approx \sum_{i=2}^{2N+1} W_i \cdot (\log(S_{ji}) - \mu_{jl}^{\mathrm{LOG}}) \cdot (\log(S_{j'i}) - \mu_{j'l}^{\mathrm{LOG}}). \tag{6.28}$$

where each vector sigma point $\mathbf{S}_i$ was obtained using Equations 6.25 and 6.26 with $\boldsymbol{\mu}_l^y = \boldsymbol{\mu}_l^{\mathrm{MEL}}$, $\boldsymbol{\Sigma}_l^y = \boldsymbol{\Sigma}_l^{\mathrm{MEL}}$ and N is equal to the number of Bark filterbanks J.

6.2.4 Propagation through the Discrete Cosine Transform

The cosine transform given in Equation 3.18 is a linear transformation and therefore can be easily computed by setting

$$T_{ij} = \sqrt{\frac{2}{J}} \cos\left(\frac{\pi i}{J}(j - 0.5)\right), \tag{6.29}$$

and using Equations 6.14 and 6.15. Despite the fact that the full covariance after the DCT transformation can be obtained and used in the recognition process (see Section 3.6), only the values of the diagonal are used in the experimental setup[6], consequently the mean and covariance can be obtained as

$$\mu_{il}^{\text{CEPS}} = \sum_{j=1}^{J} T_{ij} \mu_{jl}^{\text{LOG}} \tag{6.30}$$

and

$$\Sigma_{iil}^{\text{CEPS}} = \sum_{j=1}^{J} \sum_{j'=1}^{J} T_{ij} T_{ij'} \Sigma_{jj'l}^{\text{LOG}}. \tag{6.31}$$

The eventual use of cepstral-liftering poses no difficulty since it is also a linear transformation, that multiplies each T_{ij}.

6.3 Uncertainty Propagation for the RASTA-PLP Cepstral Features

6.3.1 Propagation through the PSD

Since a complex Gaussian distribution has a Rice distributed amplitude for which the n^{th} order moments are known, the mean of the uncertain PSD features can be computed from Equation 6.8 as

$$\mu_{kl}^{\text{PSD}} = E\{|X_{kl}|^2\} = \lambda_{kl} + |\hat{X}_{kl}|^2, \tag{6.32}$$

whereas the covariance can be computed similarly to the STSA case as

$$\Sigma_{kkl}^{\text{PSD}} = E\{|X_{kl}|^4\} - E\{|X_{kl}|^2\}^2 = 2\lambda_{kl}|\hat{X}_{kl}|^2 + \lambda_{kl}^2, \tag{6.33}$$

where, $E\{|X_{kl}|^4\}$ is obtained by solving Equation 6.4 for $n = 4$ using [Gradshteyn 2007, Eq. 8.970.5]. The third and fourth cumulants can be obtained accordingly by applying [Gradshteyn 2007, Eqs. 8.970.6, 8.970.7] to obtain the moments of sixth and eighth order, leading to the following expression for the third and fourth order cumulants

$$\text{K3}_{kl}^{\text{PSD}} = 2\lambda_{kl}^2 \left(\lambda_{kl} + 3|\hat{X}_{kl}|^2\right), \tag{6.34}$$

[6]Note that this does not imply that the correlation induced by the Mel-filterbank can be therefore ignored since it necessary for the computation of the diagonal values of the covariance as well.

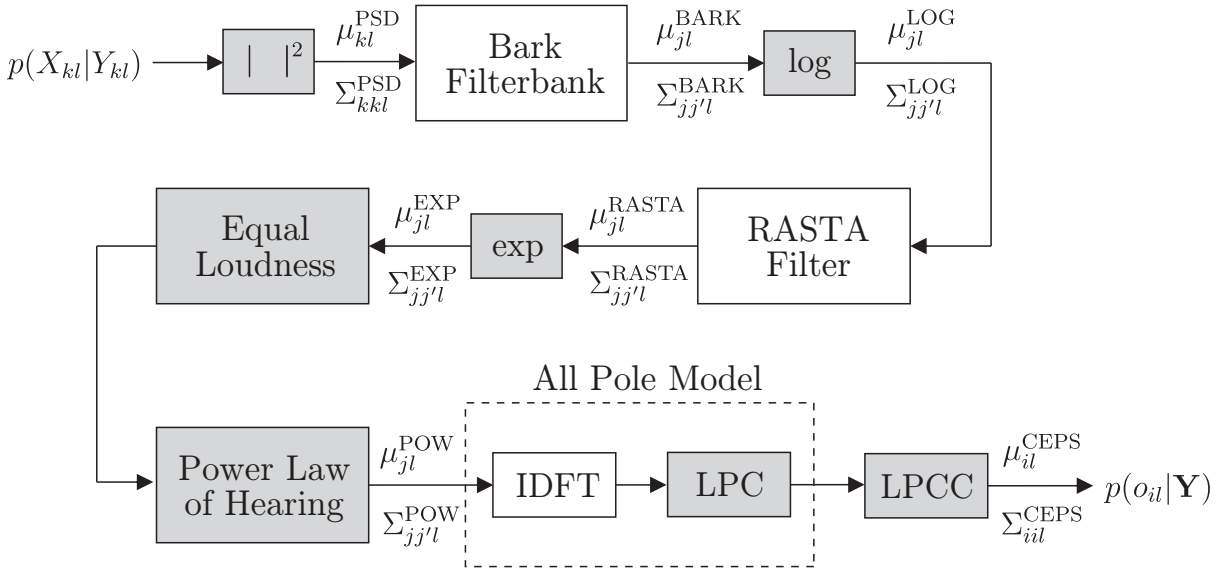

Figure 6.3: RASTA-PLP feature extraction starting in the STFT domain, extended to propagate second order information into the cepstral domain. Blocks corresponding to non-linear transformations are displayed in gray. Note that only diagonal covariances are considered in the initial (PSD) and final transformations (CEPS).

and

$$\mathrm{K4}_{kl}^{\mathrm{PSD}} = 6\lambda_{kl}^3 \left(\lambda_{kl} + 4|\hat{X}_{kl}|^2\right). \tag{6.35}$$

6.3.2 Propagation through the Bark Filterbank

As in the case of the Mel-filterbank, the Bark filterbank, defined in Equation 3.21, is a linear transformation and therefore propagating mean and covariance requires only matrix multiplications, hence giving

$$\mu_{jl}^{\mathrm{Bark}} = \sum_{k=1}^{K} V_{jk}\mu_{kl}^{\mathrm{PSD}} \tag{6.36}$$

and

$$\Sigma_{jj'l}^{\mathrm{Bark}} = \sum_{k=1}^{K}\sum_{k'=1}^{K} V_{jk}V_{j'k'}\Sigma_{kk'l}^{\mathrm{PSD}} = \sum_{k=1}^{K} V_{jk}V_{j'k}\Sigma_{kkl}^{\mathrm{PSD}} \tag{6.37}$$

where V_{jk} was defined in Section 3.3.2.

6.3.3 Propagation through the Logarithm of the Bark-PSD

During the tests carried out in this thesis, it was observed that the distribution of the Bark-PSD uncertain features is more skewed than the Mel-STSA features. A possible explanation for this phenomenon is the fact that STSA features correspond to a sum of Rice distributions of different parameters, whereas the PSD features correspond to a weighted sum of non-central Chi square distributions of different parameters.
It was also observed that the log-Normality assumption offered a good approximation for the distribution of the filterbank-PSD features regardless of which of the two filterbanks was used. The log-Normality assumption, also used in other propagation approaches [Gales 1995, Kolossa 2005, Stouten 2006b], is equivalent to the assumption of the logarithm of the uncertain Bark-PSD features being Gaussian distributed [Johnson 1970, Chap. 14]. The corresponding statistics in the Bark domain can be then computed by propagating back the Gaussian distribution from the log domain through the exponential transformation (see derivation in Section 6.7.1) and matching first and second moments, thus obtaining

$$\mu_{jl}^{\mathrm{LOG}} = \log(\mu_{jl}^{\mathrm{Bark}}) - \frac{1}{2}\log\left(\frac{\Sigma_{ijl}^{\mathrm{BARK}}}{\mu_{jl}^{\mathrm{BARK}}\mu_{jl}^{\mathrm{BARK}}} + 1\right), \tag{6.38}$$

and

$$\Sigma_{jj'l}^{\mathrm{LOG}} = \log\left(\frac{\Sigma_{jj'l}^{\mathrm{BARK}}}{\mu_{jl}^{\mathrm{Bark}}\mu_{j'l}^{\mathrm{Bark}}} + 1\right). \tag{6.39}$$

6.3.4 Propagation through RASTA Filter

As an IIR filter, the RASTA-filter is a recursive linear transformation corresponding to the difference equation

$$\mathbf{y}_l = \sum_{d=0}^{4} b_d \mathbf{x}_{l-d} - a_1 \mathbf{y}_{l-1}. \tag{6.40}$$

Here $\mathbf{y}_l$ is a column vector containing the l^{th} frame of features, $\mathbf{x}_l \cdots \mathbf{x}_{l-4}$, and $\mathbf{y}_{l-1}$ correspond to the previous logarithm domain input and RASTA domain output frames respectively. The scalars $b_0 \cdots b_4, a_1$ are the normalized feedforward and feedback coefficients. Computing the propagation of the mean through this transformation is immediate due to the linearity of the expectation operator. Applying Equation 6.14, we obtain

$$\boldsymbol{\mu}_l^{\mathrm{RASTA}} = E\{\mathbf{y}_l\} = \sum_{d=0}^{4} b_d \boldsymbol{\mu}_l^{\mathrm{LOG}} - a_1 \boldsymbol{\mu}_{l-1}^{\mathrm{RASTA}}, \tag{6.41}$$

where the last term $\boldsymbol{\mu}_{l-1}^{\mathrm{RASTA}}$ was obtained in the previous frame computation. The computation of the covariance is however more complex due to the created time-correlation between inputs and outputs and the correlation between different features created after

the filterbank. If we consider the feature covariance matrix for frame l we have that by [Chrysostomos 1993, Eq. 2.4]

$$\mathbf{\Sigma}_l^{\text{RASTA}} = E\{\mathbf{y}_l\mathbf{y}_l^T\} - E\{\mathbf{y}_l\}E\{\mathbf{y}_l\}^T, \tag{6.42}$$

where the second term can be directly calculated from Equation 6.41 but the first term must be solved. By replacing Equation 6.40 into $E\{\mathbf{y}_l\mathbf{y}_l^T\}$ we obtain

$$E\{\mathbf{y}_l\mathbf{y}_l^T\} = E\left\{\left(\sum_{d=0}^{4} b_d\mathbf{x}_{l-d} - a_1\mathbf{y}_{l-1}\right)\left(\sum_{d=0}^{4} b_d\mathbf{x}_{l-d} - a_1\mathbf{y}_{l-1}\right)^T\right\}, \tag{6.43}$$

taking into account the linearity of the expectation operator, the distributive property of matrix multiplication and the properties of the transpose operator we can further simplify this as

$$\begin{aligned} E\{\mathbf{y}_l\mathbf{y}_l^T\} = \sum_{d=0}^{4} b_d^2 E\{\mathbf{x}_{l-d}\mathbf{x}_{l-d}^T\} + a_1^2 E\{\mathbf{y}_{l-1}\mathbf{y}_{l-1}^T\} - \sum_{d=0}^{4} b_d a_1 E\{\mathbf{y}_{l-1}\mathbf{x}_{l-d}^T\} \\ - \sum_{d=0}^{4} b_d a_1 E\{\mathbf{x}_{l-d}\mathbf{y}_{l-1}^T\}, \end{aligned} \tag{6.44}$$

where the last two terms are one the transpose of the other and account for the correlation between the actual inputs $\mathbf{x}_l \cdots \mathbf{x}_{l-4}$ and the shifted output $\mathbf{y}_{l-1}$. Since this is not zero due to the action of the IIR filter, it has to be further calculated. Both terms are computed analogously. If we take for example the term $E\{\mathbf{y}_{l-1}\mathbf{x}_{l-d}^T\}$, we can further expand it by replacing in it Equation 6.40 shifted one time unit, thus leading to

$$\begin{aligned} E\{\mathbf{x}_{l-d}\mathbf{y}_{l-1}^T\} = E\left\{\mathbf{x}_{l-d}\left(\sum_{p=0}^{4} b_p\mathbf{x}_{l-1-p} - a_1\mathbf{y}_{l-2}\right)^T\right\} = \\ \sum_{p=0}^{4} b_p E\{\mathbf{x}_{l-d}\mathbf{x}_{l-1-p}^T\} - a_1 E\{\mathbf{x}_{l-d}\mathbf{y}_{l-2}^T\}. \end{aligned} \tag{6.45}$$

Since the inputs are assumed uncorrelated (in time), the first term will only be non zero if inputs are the same, that is if $d = 1 + p$, which simplifies the formula to

$$E\{\mathbf{x}_{l-d}\mathbf{y}_{l-1}^T\} = b_{d-1}E\{\mathbf{x}_{l-d}\mathbf{x}_{l-d}^T\} - a_1 E\{\mathbf{x}_{l-d}\mathbf{y}_{l-2}^T\}. \tag{6.46}$$

The second term of this formula is again an input and output correlation that can be expanded as in Equation 6.45 and simplified the same way. By recursively replacing the time shifted equivalents of Equation 6.40 into the remaining input output correlation we reach a point in which the shifted output is no longer dependent on the inputs, thus obtaining

$$E\{\mathbf{x}_{l-d}\mathbf{y}_{l-1}^T\} = b_{d-1}E\{\mathbf{x}_{l-d}\mathbf{x}_{l-d}^T\} - a_1 b_{d-2}E\{\mathbf{x}_{l-d}\mathbf{x}_{l-d}^T\} + a_1^2 E\{\mathbf{x}_{l-d}\mathbf{y}_{l-3}^T\} = \sum_{q=1}^{d}(-1)^{q-1}a_1^{q-1}b_{d-q}E\{\mathbf{x}_{l-d}\mathbf{x}_{l-d}^T\}, \quad (6.47)$$

substituting this result into Equation 6.44 and replacing this last equation as well as the propagated mean in Equation 6.41 into Equation 6.42 we obtain the following formula for the covariance

$$\boldsymbol{\Sigma}_l^{\text{RASTA}} = \sum_{d=0}^{4} b_d^2 E\{\mathbf{x}_{l-d}\mathbf{x}_{l-d}^T\} + a_1^2 E\{\mathbf{y}_{l-1}\mathbf{y}_{l-1}^T\} + 2\sum_{d=0}^{4} b_d \sum_{q=1}^{d}(-1)^q a_1^q b_{d-q} E\{\mathbf{x}_{l-d}\mathbf{x}_{l-d}^T\} - \boldsymbol{\mu}_l^{\text{RASTA}}\left(\boldsymbol{\mu}_l^{\text{RASTA}}\right)^T, \quad (6.48)$$

where $E\{\mathbf{y}_{l-1}\mathbf{y}_{l-1}^T\}$ is obtained from the previous frame computation. This closed-form solution is a particular case of uncertainty propagation through a generic IIR filter. Such a transformation can be always solved using the conventional matrix solutions in Equations 6.14 and 6.15, as long as the output-output and input-output correlations are taken into consideration and included in the input covariance matrix. The computation of the appearing correlations can be performed on the fly by solving the linear transformations at each step.

It should also be noted that the RASTA filter induces a correlation between features of different frames, $E\{\mathbf{y}_l\mathbf{y}_{l-d}^T\}$. In order to achieve full propagation of second order information, this correlation should be taken into account. However, for the sake of simplicity, the created time correlation is ignored in the upcoming uncertainty propagation steps. Experimental results show that this correlation has a low influence.

6.3.5 Propagation through Pre-Emphasis, Power Law of Hearing and Exponential

Both pre-emphasis and power law of hearing transformations, summarized in the formula

$$H_{jl} = \left(\psi_j B_{jl}\right)^{0.33}, \quad (6.49)$$

can be performed in the log domain where they can be formulated as the following linear transformation

$$H_{jl} = \exp(0.33\log(\psi_j) + 0.33\log(B_{jl})). \quad (6.50)$$

Since the assumed log-Normal distribution of uncertainty in the Bark domain led to Gaussian distributed uncertainty in the log domain, this does not change either after the RASTA nor the pre-emphasis and power law of hearing transformations. Consequently, applying the exponential transformation leads to a log-Normal distribution.

The mean and covariance of this distribution can be computed by combining the solution for the linear transformations, Equations 6.14 and 6.15, with the log-Normal forward propagation through the exponential given by Equations 6.74 and 6.75, yielding

$$\mu_{jl}^{\mathrm{EXP}} = \exp\left(0.33\log(\psi_j) + 0.33\mu_{jl}^{\mathrm{RASTA}} + \frac{0.33^2\Sigma_{jjl}^{\mathrm{RASTA}}}{2}\right), \tag{6.51}$$

and

$$\Sigma_{jj'l}^{\mathrm{EXP}} = \mu_{jl}^{\mathrm{EXP}}\mu_{j'l}^{\mathrm{EXP}}\left(\exp\left(0.33^2\Sigma_{jj'l}^{\mathrm{RASTA}}\right) - 1\right). \tag{6.52}$$

6.3.6 Propagation through the All Pole Model and LPCC Computation

As explained in Section 3.3.2, the computation of the cepstral coefficients from the all pole model implies computing the inverse Fourier transformation, implemented in this case with the inverse discrete Fourier transform (IDFT), obtaining the linear prediction coefficients with the Levinson-Durbin recursion and finally computing the cepstra. Although some of these steps, like the inverse DFT, are linear, the most suitable solution here is to use the unscented transform to solve all these transformations in one single non-linear transformation $\mathbf{g}$. This is so because the dimensionality of the features is low enough and experimental results show that the asymmetry of the PDF is relatively low. The computation of the LPCC coefficients mean and covariance can be obtained using the corresponding unscented transform formulas in Equations 6.23 and 6.24 with $\boldsymbol{\mu}_l^y = \boldsymbol{\mu}_l^{\mathrm{EXP}}$, $\boldsymbol{\Sigma}_l^y = \boldsymbol{\Sigma}_l^{\mathrm{EXP}}$

6.4 Uncertainty Propagation for the ETSI ES 202 050 Standard

To provide an example of a real world implementation of the piecewise uncertainty propagation, the propagation problem was solved for the advanced front-end of the ETSI standard, which was introduced in Section 3.4. This standard uses a Mel-cepstrum based feature extraction with PSD features instead of the STSA features. Mel-PSD features do however behave identically to the Bark-PSD features explained in Section 3.3.2 and thus the uncertainty propagation approach explained in this section remains perfectly valid. Figure 6.4 depicts the modifications performed in the advanced front-end with respect to the original displayed in Figure 3.6. These modifications include both the estimation of uncertainty, explained in Section 5.2.1, and uncertainty propagation. Propagation through the cepstrum calculation block is obtained by applying Equations 6.32, 6.33, 6.36, 6.37, 6.38, 6.39, 6.30 and 6.31, taking into account the change between Bark and Mel matrices. Propagation through the log-energy, blind equalization and server side computations will be explained in the next sections.

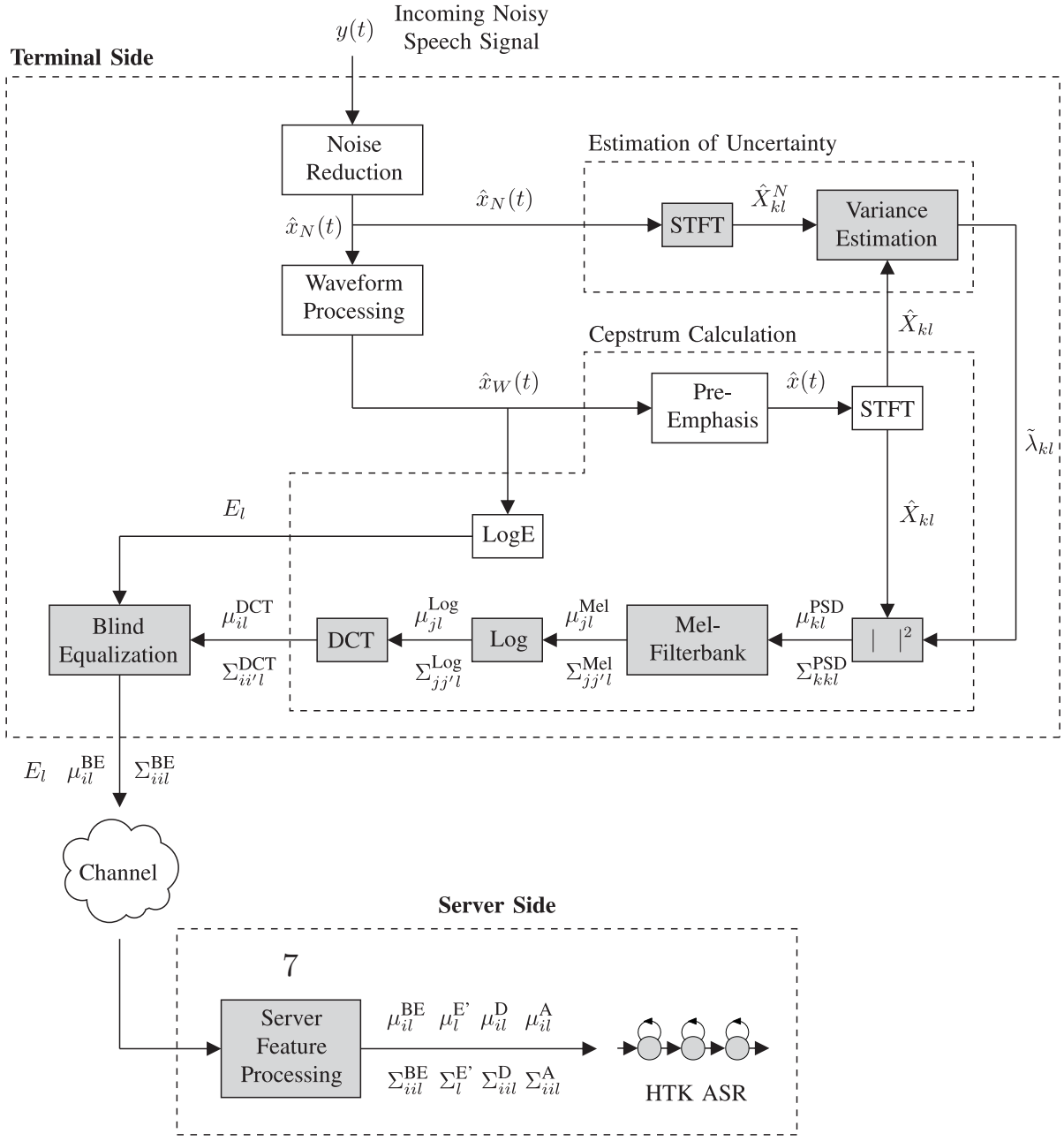

Figure 6.4: Extension of the ETSI ES 202 050 standard to support uncertainty propagation. Steps added or modified to allow uncertainty propagation are displayed in gray. Blocks corresponding to coding, transmission and decoding have been omitted.

6.4.1 Propagation of the Complex Gaussian Model Into the Time Domain

Given the uncertain Fourier spectrum of a signal $\mathbf{X}$, where each X_{kl} is statistically independent of the others and complex Gaussian distributed according to Equation 6.1, it is possible to invert $\mathbf{X}$ back to the time domain with the inverse short-time Fourier transform ISTFT, thus obtaining a Gaussian distributed uncertain time domain signal $x(t)$. In order to do this, it is first necessary to apply the IDFT to obtain, from each frame of uncertain Fourier coefficients $X_{1l} \dots X_{Kl}$, the corresponding time-domain frame $x_{1l} \dots x_{Nl}$, where K denotes the number of bins up to half the sampling frequency and $N = 2K - 2$ the size of the frame[7]. In order apply the IDFT, it is first necessary to obtain the whole spectrum (N bins), including the frequency bins over half the sampling frequency. This can be obtained as

$$\mathbf{X}'_l = [X_{1l} \dots X_{Kl}, X^*_{K-1l} \dots X^*_{2l}] \tag{6.53}$$

where X^*_{kl} denotes the conjugate of X_{kl}. Since the complex conjugate only implies multiplying the imaginary component of the Fourier coefficient by -1, we have that

$$E\{X^*_{kl}\} = E\{X_{kl}\}^*, \tag{6.54}$$

whereas the variance of the the conjugate remains unchanged. Taking this into account, the mean of the whole spectrum frame corresponds to

$$E\{\mathbf{X}'_l\} = \hat{\mathbf{X}}'_l = [\hat{X}_{1l} \dots \hat{X}_{Kl}, \hat{X}^*_{K-1l} \dots \hat{X}^*_{2l}]. \tag{6.55}$$

The correlations between real and imaginary components of the complex valued coefficients can be obtained using the definition of complex Gaussian distribution (see Section 2.3.1). If we consider the real X^R_{kl} and imaginary X^I_{kl} components of the spectrum separately, we have that the resulting diagonals of their correlation matrices yield

$$E\{X'^R_{kl} X'^R_{kl}\} = \mathrm{Var}\{X'^R_{kl}\} + E\{X'^R_{kl}\}^2 = \lambda_{kl}/2 + \left(\hat{X}^R_{kl}\right)^2, \tag{6.56}$$

and

$$E\{X'^I_{kl} X'^I_{kl}\} = \mathrm{Var}\{X'^I_{kl}\} + E\{X'^I_{kl}\}^2 = \lambda_{kl}/2 + \left(\hat{X}^I_{kl}\right)^2, \tag{6.57}$$

for $k \in [1 \cdots N]$. In addition to these correlations, the appearing correlations between the original Fourier coefficients and their conjugates (excluding the DC component) have to be considered. These correspond to

[7] In the particular case of the ETSI standard STFT, $K = 129$, $N = 256$. Note also that the ETSI STFT uses zero padding. Since this is a trivial operation it is here omitted for clarity.

$$E\{X'^R_{kl}X'^R_{k'l}\} = \lambda_{kl}/2 + \left(\hat{X}^R_{kl}\right)^2, \tag{6.58}$$

and

$$E\{X'^I_{kl}X'^I_{k'l}\} = -\lambda_{kl}/2 - \left(\hat{X}^I_{kl}\right)^2, \tag{6.59}$$

with $k \in [2 \cdots N-1]$ and $k' = N - k + 1$. Taking this into account, we consider the IDFT transformation defined as

$$x_{t'l} = \mathrm{Re}\left\{\frac{1}{N}\sum_{k=1}^{N} X'_{kl} \exp\left(2\pi i \frac{(t'-1)(k-1)}{N}\right)\right\}, \tag{6.60}$$

where $t' \in [1 \cdots N]$ is the sample index within the frame.
This transformation can be expressed as the following matrix operation[8]

$$\mathbf{x}_l = \mathrm{Re}\left\{\mathbf{F}\mathbf{X}_l^T\right\}, \tag{6.61}$$

where i is the imaginary unit and

$$F_{kt'} = F^R_{kt'} + iF^I_{kt'} = \frac{1}{N}\exp\left(2\pi i \frac{(t'-1)(k-1)}{N}\right), \tag{6.62}$$

is the inverse Fourier matrix expressed in cartesian and polar coordinates. This last term might also include the inverse window function if one is used.
Since X'_{kl} can be decomposed into its real and imaginary components

$$X'_{kl} = X'^R_{kl} + iX'^I_{kl}, \tag{6.63}$$

by using the properties of the complex numbers, Equation 6.61 can be expressed as the following linear transformation involving real valued matrices

$$\mathbf{x}_l = \mathrm{Re}\left\{\mathbf{F}\mathbf{X}_l^T\right\} = \mathbf{F}^R\left(\mathbf{X}'^R_l\right)^T - \mathbf{F}^I\left(\mathbf{X}'^I_l\right)^T \tag{6.64}$$

where $F^R_{kt}, F^I_{kt} \in \mathbb{R}$ and by the definition of complex Gaussian distribution, $\mathbf{X}'^R_l$, $\mathbf{X}'^I_l$ are multivariate Gaussian distributed with the correlations given by Equations 6.56, 6.57, 6.58 and 6.59. Consequently $\mathbf{x}_l$ will be multivariate Gaussian distributed with mean

$$E\{\mathbf{x}_l\} = \mathbf{F}^R\left(\hat{\mathbf{X}}'^R_l\right)^T - \mathbf{F}^I\left(\hat{\mathbf{X}}'^I_l\right)^T \tag{6.65}$$

[8] Re{} operates here over the vector elementwise.

and correlation

$$E\{\mathbf{x}_l\mathbf{x}_l^T\} = \mathbf{F}^R E\left\{\mathbf{X}_l'^R(\mathbf{X}_l'^R)^T\right\}\left(\mathbf{F}^R\right)^T + \mathbf{F}^I E\left\{\mathbf{X}_l'^I(\mathbf{X}_l'^I)^T\right\}\left(\mathbf{F}^I\right)^T, \tag{6.66}$$

where for this last formulation the fact that real and imaginary components are statistically independent has been used, and the correlation matrices in Equations 6.56, 6.57, 6.58 and 6.59 have been expressed in matrix form.
Once each frame of the uncertain Fourier spectrum $\mathbf{X}_l$ has been transformed into an uncertain time domain frame $\mathbf{x}_l$, the uncertain time domain signal can be recovered using the overlap and add method (see [Cohen 2001, Eq. 3]).

6.4.2 Propagation through Log-Energy Transformation

The computation of the log-energy is given by the formula (Equation 3.28)

$$E_l = \log\left(\sum_{t'=1}^{N} x_{t'l}^2\right), \tag{6.67}$$

where $x_{t'l}$ corresponds to each of the time domain frame elements. As shown in the last section, for the used uncertainty model, each frame of time-domain elements $x_{1l}\cdots x_{Nl}$ is multivariate Gaussian distributed with a covariance matrix given by Equation 6.66. Consequently, the mean of the squared time domain elements $x_{1l}^2\cdots x_{Nl}^2$ will correspond to the diagonal of Equation 6.66. Furthermore, given the formula which relates the fourth order cross cumulant and the raw moments of a random variable [Chrysostomos 1993, Eq. 2.7] and provided that a multivariate Gaussian distribution has a fourth order cross cumulant equal to zero, we have

$$E\{x_{t'l}^2x_{t''l}^2\} = E\{x_{t'l}^2\}E\{x_{t''l}^2\} + 2E\{x_{t'l}x_{t''l}\}^2 - 2E\{x_{t'l}\}^2E\{x_{t''l}\}^2, \tag{6.68}$$

which enables us to compute the full covariance of the squared Gaussian variables from the mean and covariance given by Equations 6.65 and 6.66. To complete the transformation in Equation 3.28, the logarithm of the sum of $x_{1l}^2\cdots x_{Nl}^2$ has to be computed. The sum is a linear transformation and thus can be computed with Equations 6.14 and 6.15, whereas the logarithm has to be approximated here by using the log-Normal assumption specified by Equations 6.38 and 6.39. Combining the formulas for both mean and covariance results in

$$\Sigma_l^{\mathrm{LogE}} = \log\left(\sum_{t'=1}^{N}\sum_{t''=1}^{N} E\{x_{t'l}^2x_{t''l}^2\}\right) - 2\log\left(\sum_{t'=1}^{N} E\{x_{t'l}^2\}\right) \tag{6.69}$$

and

$$\mu_l^{\mathrm{LogE}} = \log\left(\sum_{t'=1}^{N} E\{x_{t'l}^2\}\right) - \frac{1}{2}\Sigma_l^{\mathrm{LogE}}, \tag{6.70}$$

where in this case, for the sake of simplicity, Equation 6.39 has been re-expressed in terms of mean and correlation instead of mean and covariance using [Chrysostomos 1993, Eq. 2.4].

6.4.3 Propagation through Blind Equalization

The propagation of uncertainty through the blind equalization transformation, given by Equations 3.29, 3.30 and 3.31, can be rather complicated if we consider that it includes the product of two random variables, the step-factor μ_l, obtained from a non-linear transformation of the log-energy of the previous frame E_{l-1} (see Section 6.4.2 and Equation 3.31) and the Gaussian distributed cepstral coefficient C_{il}. As it will be shown in the experimental setup, this propagation problem can be simplified by considering the log-energy as deterministic and using the mean of the log energy $\mu_{l-1}^{\mathrm{LogE}}$ instead of E_{l-1}. In this case the propagation problem can be decoupled into two conventional linear propagation problems solvable by using Equations 6.14 and 6.15 on Equations 3.29 and 3.30 individually. The correlations induced between features of different frames are also ignored here.

6.4.4 Transmission and Server Side Computations

As explained in Section 3.4, the transmission of the uncertain features through the channel requires a previous quantization. This quantization was handled here ideally in the sense that the means were quantized with the available codebook and the variances ideally transmitted to the server side.
It should also be noted that the use of full covariances for uncertainty propagation does not imply that the full matrix has to be propagated. As long as only the diagonals of the covariance matrices are used for the recognition process, full covariances are only needed at the terminal side. Server side operations do not involve joint transformations between features of the same frame and there is no additional loss of accuracy when full covariances are not available.
The computation of delta and acceleration coefficients (see Section 3.3.5), which also takes place on the server side [ETSI 2007, sec. 9.2], is a trivial linear transformation. As the final step on the server side, a second voice activity detection system is used to eliminate frames containing only noise [ETSI 2007, sec. 9.3]. Since the logic of this system is relatively complex, the mean of the uncertain features was used for this purpose.

6.5 Taking Full Covariances Into Consideration

6.5.1 Covariances Between Features of the Same Frame

As commented at the beginning of this chapter, one of the main limiting factors when using uncertainty propagation for robust ASR is the consumption of computation resources. An ASR system can be implemented in multiple types of devices, from embedded systems to distributed systems spanning across several servers. Such systems have different costs for the different operations involved in the uncertainty propagation

algorithms and therefore determining the exact computational cost is not possible.
One easy modification which provided a sensible reduction of computational time was ignoring the correlation created between the features after the filterbank transformations (see Sections 6.2.2 and 6.3.2). The assumption of diagonal covariances after the filterbank reduces the number of operations needed for the computation of the covariances in most transformations. In the case of the Mel, DCT, Bark, log-Bark, RASTA, exponential and power law of hearing transformations, the number of operations is divided by the dimension of the Mel/Bark features (number of filters) J[9]. Furthermore it also greatly simplifies the unscented transform algorithm (see Section 6.2.3.1) since, apart from reducing the number of computations, the matrix square root needed for the determination of the sigma points turns into a conventional square root.
The assumption of diagonal covariance can not only be applied after the Mel/Bark filterbanks, but also after the IDFT transformation when computing the log-energy (see Section 6.4.2).
The effect of both assumptions will be analyzed in the experimental setup in Chapter 7.

6.5.2 Covariance Between Features of Different Frames

Apart from the RASTA filter, various transformations where presented in Chapter 3 that imply combining features of different time-frames. The delta, acceleration and cepstral mean subtraction transformations given by Equations 3.26 and 3.27 can be described as FIR filters (see [Proakis 2007, Sec. 2.3.7]), are thus linear, and can be solved by applying Equations 6.14 and 6.15. Nevertheless, if a previous transformation induced a correlation between features of different frames (i.e. RASTA filter), this should be included in the covariance matrix. For the typical output of RASTA filtering from Log-Bark-PSD features, for example, the uncertain feature distribution is described by a $J \cdot L$ element mean matrix and L intra-frame covariance matrices of size J^2. If all inter-frame time correlations are to be considered, a $(L \cdot J)^2$ matrix is needed. This matrix can be optimized taking into account that the time correlations only span between neighboring frames but it remains nevertheless too big an increase in computational costs. Inter-frame correlation is therefore ignored in delta, acceleration and cepstral mean subtraction transformations. As it will be seen in the experimental setup in Chapter 7, this does not affect the accuracy of the uncertainty propagation algorithms much.

6.6 Summary Table of Transformations and Assumed Distributions

As a summary of this chapter, the following table contains all the transformations, for which a solution for uncertainty propagation was found, along with the assumptions performed and statistical information available.

[9]The value of J varies with the implementations but is usually around 20.

STFT Uncertainty Model (See section Section 5.1.1)			
Transformation	Assumed Distribution	Computed	Validity
1. Fourier Coefficient	Complex Gaussian	PDF	Premise
MFCC (See Section 6.2)			
Transformation	Assumed Distribution	Computed	Validity
2. STSA	Rice	PDF	Follows automatically from 1.
3. Mel/Bark-STSA	Symmetric	n^{th} Cum.	Symmetry has to be assessed.
4. Log-Mel/Bark-STSA	None	Mean*, Cov.*	Accuracy unscented transform.
5. DCT	Gaussian	PDF*	Gaussianity has to be assessed.
RASTA-PLP (See Section 6.3)			
Transformation	Assumed Distribution	Computed	Validity
6. PSD	Scaled non-central Chi Square	n^{th} Cum.	Follows automatically from 1.
7. Mel/Bark-PSD	Log-Normal	PDF*/n^{th} Cum.	Log-Normality has to be assessed
8. Log-Mel/Bark-PSD	Gaussian	PDF*	Follows automatically from 7.
9. RASTA IIR Filter	Gaussian	PDF*	Follows automatically from 7.
10. Power Law of Hearing Filter	Gaussian	PDF*	Follows automatically from 7.
11. Exponential	Log-Normal	PDF*	Follows automatically from 7.
12. LPCC	Gaussian	PDF*	Gaussianity has to be assessed.
Other (See Sections 6.5.2 and 6.4)			
Transformation	Assumed Distribution	Computed	Validity
13. FIR Filters after (5, 12, 16)	Gaussian	PDF$^+$	if initial distribution is Gaussian.
14. IDFT	Gaussian	PDF	Follows automatically from 1.
15. Log-energy	Gaussian	PDF*	Gaussianity has to be assessed.
16. Blind-Equalization	Log-energy is deterministic	PDF*	Gaussianity has to be assessed.

Table 6.1: Assumed distributions after the different steps of feature extraction. (*) Approximated PDF or parameters. (+) Time correlations ignored.

6.7 Derivations

6.7.1 Propagation through the Exponential

Let $\mathbf{L}_l$ be a multivariate Gaussian distribution of vector mean $\boldsymbol{\mu}_l^{\mathrm{LOG}}$ and covariance matrix $\boldsymbol{\Sigma}_l^{\mathrm{LOG}}$ representing the uncertainty distribution in frame l of the logarithm domain features. The propagation of its first moment back into the Bark domain using the exponential can be attained as

$$\boldsymbol{\mu}_l^{\mathrm{Bark}} = E\{\exp(\mathbf{L}_l)\} = \int \frac{\exp(\mathbf{L}_l)}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}_l^{\mathrm{LOG}}|}} \exp\left(-\frac{1}{2}(\mathbf{L}_l - \boldsymbol{\mu}_l^{\mathrm{LOG}})\boldsymbol{\Sigma}_l^{\mathrm{LOG}}(\mathbf{L}_l - \boldsymbol{\mu}_l^{\mathrm{LOG}})^T\right) d\mathbf{L}_l, \tag{6.71}$$

whereas the propagation of the covariance can be realized by using the relation between moments and cumulants in [Chrysostomos 1993, Eq. 2.4]

$$\boldsymbol{\Sigma}_l^{\mathrm{Bark}} = E\{\exp(\mathbf{L}_l)\,(\exp(\mathbf{L}_l))^T\} - \boldsymbol{\mu}_l^{\mathrm{BARK}}\left(\boldsymbol{\mu}_l^{\mathrm{BARK}}\right)^T, \tag{6.72}$$

with

$$E\{\exp(\mathbf{L}_l)\,(\exp(\mathbf{L}_l))^T\} = \int \frac{\exp(\mathbf{L}_l)\,(\exp(\mathbf{L}_l))^T}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}_l^{\mathrm{LOG}}|}} \exp\left(-\frac{1}{2}(\mathbf{L}_l - \boldsymbol{\mu}_l^{\mathrm{LOG}})\boldsymbol{\Sigma}_l^{\mathrm{LOG}}(\mathbf{L}_l - \boldsymbol{\mu}_l^{\mathrm{LOG}})^T\right) d\mathbf{L}_l. \tag{6.73}$$

Both these equations can be solved by completing the squares in the numerator of the exponential, yielding [Gales 1995, Ap. C]

$$\mu_{jl}^{\mathrm{Bark}} = \exp\left(\mu_{jl}^{\mathrm{LOG}} + \frac{\Sigma_{jjl}^{\mathrm{LOG}}}{2}\right), \tag{6.74}$$

and

$$\Sigma_{jj'l}^{\mathrm{Bark}} = \mu_{jl}^{\mathrm{Bark}}\mu_{j'l}^{\mathrm{Bark}}\left(\exp\left(\Sigma_{jj'l}^{\mathrm{LOG}}\right) - 1\right). \tag{6.75}$$

Inverting the relations, we obtain the desired formulas in Equations 6.38 and 6.39.

7 Experimental Results

7.1 Montecarlo Simulation of Uncertainty Propagation

The uncertainty propagation algorithms proposed in Chapter 6 include various approximations and assumptions regarding the distribution of the uncertain features. In order to test these assumptions as well as the accuracy of the algorithms, various experiments were carried out with the help of Montecarlo simulations. For the complex Gaussian Fourier uncertainty model proposed in Section 5.1.1, samples of the distribution can be easily computed with a random number generator. However, since the feature extraction process often implies performing operations jointly on features of the same frame (i.e. Mel-filterbank, DCT) or on elements of different time frames (i.e. RASTA filter, blind equalization), all uncertain Fourier coefficients have to be considered at once for each signal.
To deal with this issue, the following Montecarlo simulation was implemented

1. Given a noisy signal $y(t)$ and its STFT $\mathbf{Y}$, compute an estimation of its clean STFT $\hat{\mathbf{X}}$ using any of the methods introduced in Chapters 2 and 4.

2. For the resulting matrix $\hat{\mathbf{X}}$ of size $K \cdot L$ (see Equations 2.2 and 2.3), compute the associated variance of estimation matrix $\tilde{\boldsymbol{\lambda}}$ of identical size using any of the methods introduced in Chapter 5.

3. For each estimated clean Fourier coefficient of the STFT $\hat{X}_{kl}$ and associated variance $\tilde{\lambda}_{kl}$, draw N samples of the complex Gaussian distribution given by Equation 5.1 using its relation with the Gaussian distribution (see Equation 2.61).

4. As the result of the last step, N matrix samples of size $K \cdot L$ of the uncertain STFT have been created (see Figure 7.1). Transform each of these matrix samples individually through the conventional feature extraction process, as if it were the conventional STFT of a signal.

5. For each step of the feature extraction process, use the transformed matrix samples to compute mean and covariance as well as higher order cumulants for each feature of each frame. Accumulate samples of selected features or feature statistics to generate histograms.

In order to test the validity and efficiency of the proposed uncertainty propagation algorithms, this Montecarlo simulation should be carried out with all the plausible values of $\hat{X}$ and $\tilde{\lambda}$ to be found under operating conditions. This is however impossible, nor is possible to accurately determine the relative frequency with which a finite set of

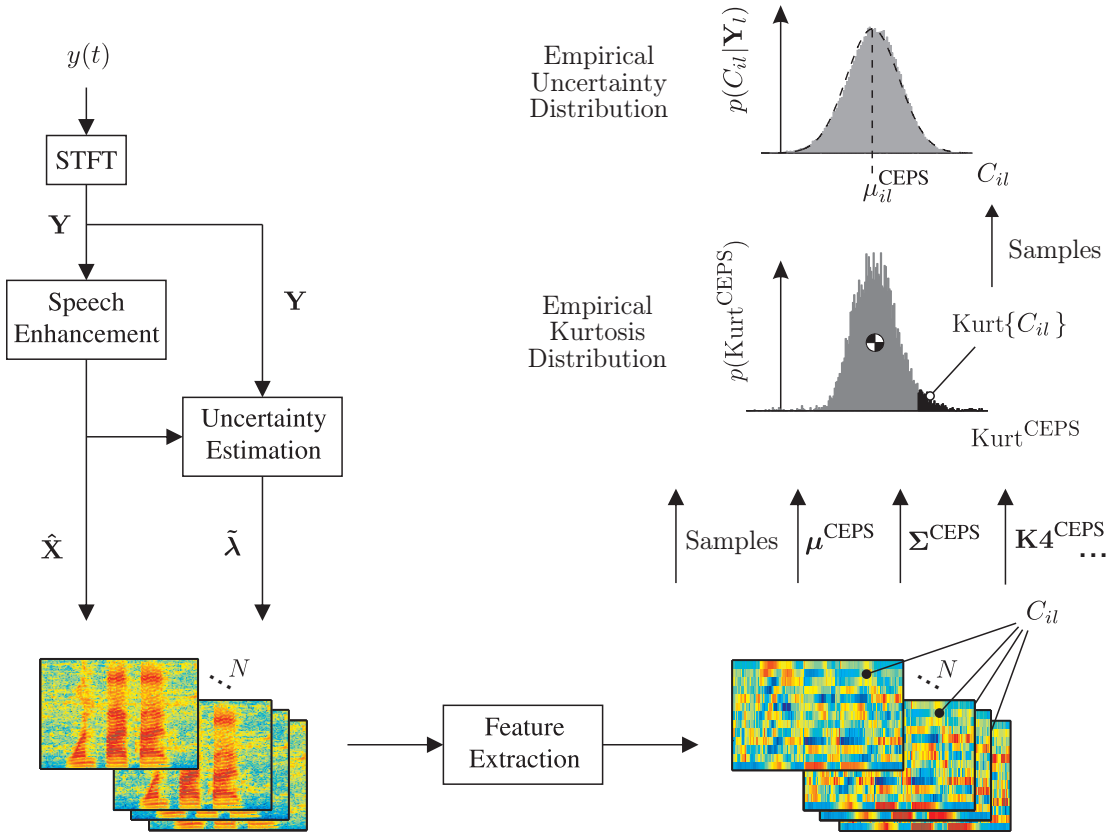

Figure 7.1: Example of the Montecarlo simulation used in the experimental setup. N matrix samples of the uncertain STFT are transformed through the studied feature extraction. Statistics like mean, covariance, and higher order cumulants are computed from the transformed samples. Histograms of a given statistic (i.e Kurtosis) can be computed using the whole matrix samples. Distributions of uncertainty for feature elements (i.e. a Mel-cepstral coefficient C_{il}) can also be computed using individual elements of each matrix sample.

operating conditions may arise. The simulation values for $\hat{X}$ and $\tilde{\lambda}$ were therefore selected based on the assumption that the parameter with the highest influence on the efficiency of uncertainty propagation is the signal to noise ratio of each noisy Fourier coefficient Y_{kl}. Other factors like speech or noise specific characteristics (speaker type, underlying phonetic transcription), speech enhancement method or uncertainty variance estimator were assumed to have no influence.

For an instationary noise signal with energy spread widely across the spectrum, added at different SNRs to a clean signal, it is very likely that all the plausible combinations of local SNRs of $\mathbf{Y}$ are attained. Consequently, a single noisy signal was created for the Montecarlo simulations by adding an instationary noise sample to a clean file of the AURORA5 database [Hirsch 2007]. The file selected was the 'crowded street' noise sample from the JEIDA database [Itahashi 1990], which has its energy spread across the

whole spectrum. The SNRs selected were -10, 0, 10 and ∞. The scaling factor for each SNR was obtained from the framed signals with no pre-emphasis using the segmental signal to noise ratio [Quackenbush 1988]. Given the clean framed time domain signal (see Equation 2.2) and scaled noise framed time domain signal x_{tl} and $\alpha \cdot d_{tl}$, respectively, the segmental SNR is defined as

$$\mathrm{SNR} = \frac{10}{L} \sum_{l=1}^{L} \log_{10} \left(\frac{\sum_{t'=1}^{N} |x_{t'l}|^2}{\sum_{t'=1}^{N} |\alpha \cdot d_{t'l}|^2} \right), \tag{7.1}$$

where L is the total number of frames. This leads to the following value for the scaling factor

$$\log_{10} \alpha = \frac{1}{20} \left(\frac{10}{L} \sum_{l=1}^{L} \log_{10} \left(\frac{\sum_{t'=1}^{N} |x_{t'l}|^2}{\sum_{t'=1}^{N} |d_{t'l}|^2} \right) - \mathrm{SNR} \right). \tag{7.2}$$

7.1.1 Assessing the Assumed PDFs through Histograms

Throughout Chapter 6, various assumptions were made regarding the distribution of uncertainty at different steps of the feature extraction process. These are summarized in Table. 6.1. In order to check the validity of the proposed uncertainty propagation algorithms, it is necessary to verify the following assumptions

1. In the MFCC features (see Section 6.2), the distribution of the features after applying the Mel-filterbank to the STSA (Mel-STSA) is assumed to be fairly symmetric. This allows the use of the unscented transform (see Section 6.2.3). Ideally this distribution should be Gaussian.

2. The final step of the MFCC features is the DCT transformation (see Section 6.2.4). The distribution of the features after applying this transformation is assumed to be Gaussian. This allows the use of modified imputation and uncertainty decoding (see Section 3.6).

3. In the RASTA-LPCC features (see Section 6.3), the distribution of the features after applying the Bark-filterbank to the PSD (Bark-PSD) is assumed to be log-Normal, or equivalently the Log-Bark-PSD features must be Gaussian distributed (see Section 6.3.3).

4. The final step of the LPCC features is the computation of the cepstra from the LPCs (see Section 6.3.6). The distribution of the features after applying this transformation is assumed to be Gaussian, which allows the use of modified imputation and uncertainty decoding.

5. In the ETSI standard features, the approximate solution for the propagation of uncertainty through the blind equalization transformation (BE) is assumed to result in Gaussian distributed features (see Section 6.4.3). This is needed for the same reasons as in points 2 and 4.

6. In the ETSI standard features (see Section 6.4), the distribution of the log-energy (LogE) is assumed to be Gaussian distributed (see Section 6.4.2). This is needed for the same reasons as in points 2, 4 and 5.

Each of these hypotheses is satisfied if the assumption of Gaussianity is met. A simple measure of non-Gaussianity often used is the kurtosis, which is zero for a Gaussian distributed random variable. The kurtosis can be computed as

$$\text{Kurt} = \frac{K4}{K2^2}, \tag{7.3}$$

where $K2$ and $K4$ correspond to the second and fourth order cumulants. For an uncertain (random) feature x, the cumulant of p^{th} order can be expressed as function of the raw moments up to order p, $E\{x\} \cdots E\{x^p\}$ using [Chrysostomos 1993, Eq. 2.4]. To compute the kurtosis, the moments for $p \leq 4$ are then needed. Given a set of N samples $x_1 \cdots x_N$ of x, obtained through the Montecarlo simulation, the raw moments were estimated using arithmetical averages of x^p as

$$E\{x^p\} \approx \overline{x^p} = \frac{1}{N} \sum_{n=1}^{N} x_n^p, \tag{7.4}$$

which due to the law of large numbers converge to the true value of the moment. Since the memory requirements were too high when storing all samples, the averages were computed in an online form as

$$\overline{x^p}_{n+1} = \frac{n\overline{x^p}_n + x_n^p}{n+1}. \tag{7.5}$$

For this experiment, a Montecarlo simulation with $N = 10^6$ matrix samples was carried out for the 4 levels of SNR (-10, 0, 10 and ∞) as previously described (see Figure 7.1). For each feature of each frame the raw moments up to order 4 and the corresponding kurtosis were computed using Equations 7.3 and 7.5. The obtained kurtosis values for the different SNRs tested were used to approximate an empirical distribution of the kurtosis for the Mel-STSA, DCT, Log-Bark-PSD, LPCC, LogE and BE features using histograms. With help of those histograms two cases where differentiated

1. Average case, corresponding to the center of probabilistic density of the empirically obtained kurtosis distribution.

2. Most unfavorable case, corresponding to the kurtosis values that depart most from the Gaussian hypothesis (zero). The 5% probabilty mass region with the most disfavorable kurtosis values was selected for this matter.

To complete this analysis, particular uncertain feature distributions with kurtosis values representative of the average and most unfavorable cases where selected. The obtained histograms were compared with the hypothesized distributions determined by using the uncertainty propagation algorithms described in Chapter 6. The parameters of each PDF where determined from the first and second order moments obtained from propagation with full covariances. To construct the histograms a $N = 10^5$ samples Montecarlo simulation was used.

Table 7.1: Average kurtosis values and value marking the limit of the 5% most unfavorable probability density region of the empirically obtained distributions.

Kurtosis	Mel-STSA	DCT	Log-Bark-PSD	LPCC	BE	LogE
AverageValue	-0.01	0.03	0.03	0.01	0.02	-0.01
Limit of the 5% region	-0.06	0.1	0.23	0.04	0.07	0.36

7.1.1.1 Results

The empirical distributions of the kurtosis are shown in Figure 7.2. The area of the histogram colored in black in Figure 7.2 represents the 5% probability density area of most unfavorable kurtosis values previously commented. The circle denotes the center of probabilic density or average case of the empirically obtained distribution. The average kurtosis values as well as the values marking the start of the 5% probability mass area of most disfavorable values are displayed in Table 7.1

From the results it is clear that all the different uncertain feature distributions remain close to zero kurtosis on average. Only the Log-Bark-PSD and LogE features exhibit values outside the 5% most unfavorable region greater than 0.1. Such deviations however do not seem to have a very strong effect on the accuracy of the propagation algorithms as will be seen in the next section. One possible reason for this is that a high percentage of the kurtosis samples with values far from zero correspond to samples with a very low variance. Consequently the error commited in the computation of the variance and mean estimation does not affect much posterior calculations. Regarding the LogE features, it has to be taken into account that, since these features have dimension one, not too many features were available to construct the histograms[1]. Kurtosis values obtained in this experiment for the LogE features might be less accurate than for the rest of the features. It is also worth noting that the empirical distributions of kurtosis obtained for each different SNR level did not differ much from the aggregated histograms which are shown in Figure 7.2. The only exception to this were the outliers appearing at high SNRs due to variances reaching values very near to zero. However, these quasi-deterministic features can be easily pruned by imposing a variance floor.

The representative cases of the distribution of uncertainty for individual features are displayed in Figures 7.3, 7.4 and 7.5. The Bark-PSD features are displayed rather than Log-Bark-PSD features since the fit between histogram and PDF can be better appreciated. Apart from the Mel-STSA, DCT, Bark-PSD, LPCC, LogE and BE features, the STSA and PSD features where also included for comparison. The assumed distributions for these two features are Rice and Non-Central Chi-Square (after normalization, see Section 5.1.2) respectively. Such distributions follow immediately from the assumed complex Gaussian uncertainty model and are thus an exact description of the uncertain features up to the accuracy of the Montecarlo simulation. Consequently, rather than average and unfavorable cases, cases with small and large Rice ratio r_{kj}, as defined in

[1]Note that for a typical experiment using the STFT of one file, N spectrum samples and 4 SNR levels, $4 \cdot K \cdot L$ Mel-STSA kurtosis samples are obtained (by averaging over the N spectrum samples), $4 \cdot I \cdot L$ LPCC kurtosis samples are obtained, whereas only $4 \cdot L$ LogE kurtosis samples are obtained. Typical values of feature dimensions are $K = 129$, $I = 13$.

Equation 2.77, where displayed. The convergence of the Rice distribution to the Gaussian distribution for large Rice ratios, commented at the end of Section 2.3.2, can be here clearly appreciated (see Figure 7.3, top, right).
Regarding the fit of the hypothesized PDFs and the empirically obtained histograms, the relatively accurate fit is coherent with the small kurtosis values seen in Figure 7.2. The only exception are the Mel-STSA features (Figure 7.3, middle) for which the average case shows a fit comparable or even worse with that of the most unfavorable case. This is probably due to the fact that skewness plays for this particular feature an important role in the non-Gaussianity of the features. Other measures of non-Gaussianity including both skewness and kurtosis could improve the identification of average and most unfavorable cases for the Mel-STSA features.

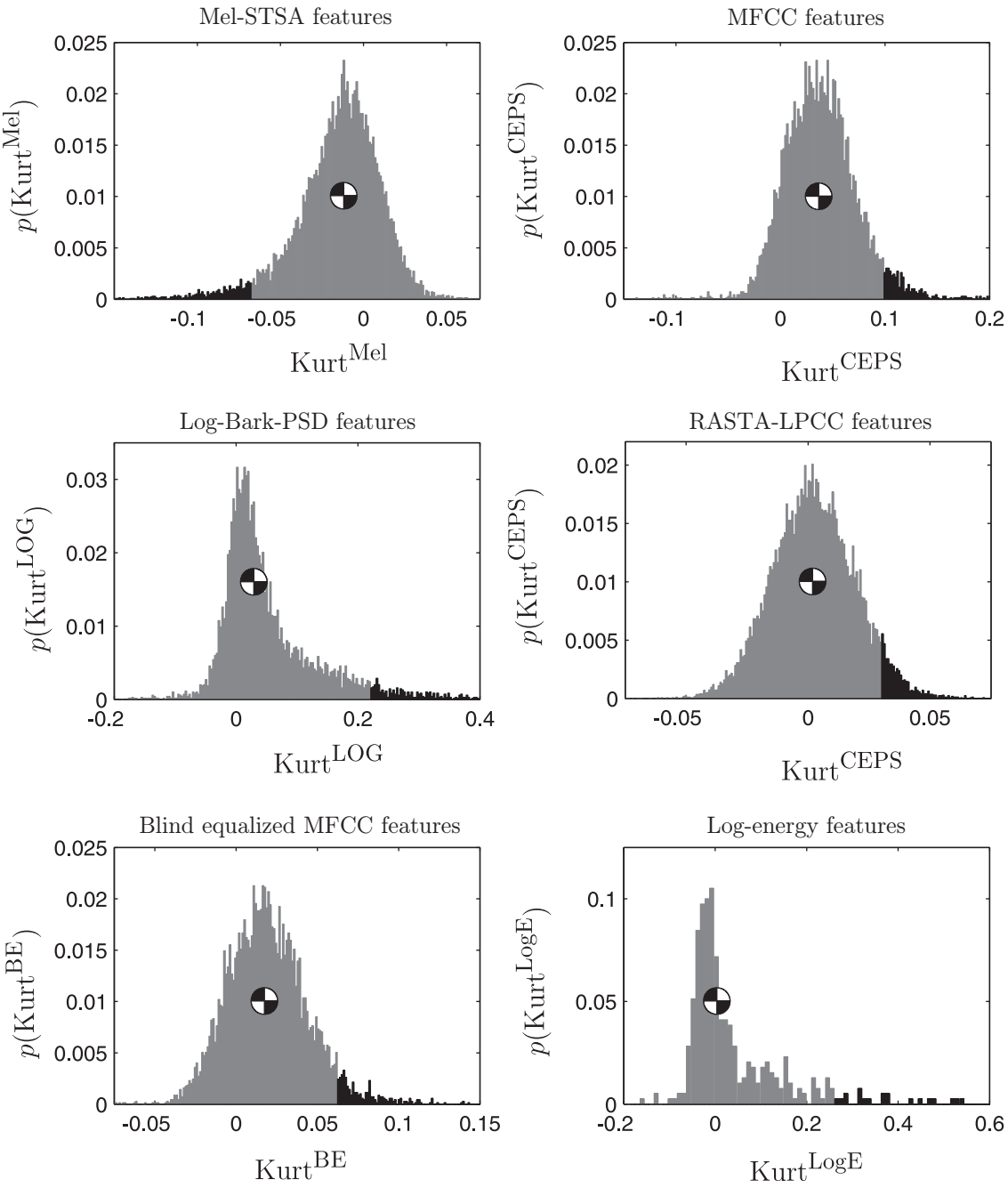

Figure 7.2: **From top-left** to **bottom-right:** Empirically obtained distribution of the kurtosis of the Mel-STSA, DCT (MFCC), Log-Bark-PSD, LPCC, BE and LogE features. The region colored in black represents the 5% probability density area with the kurtosis values that depart most from the Gaussian hypothesis whereas the circle denotes the center of probabilistic density of the empirically obtained distribution or average case.

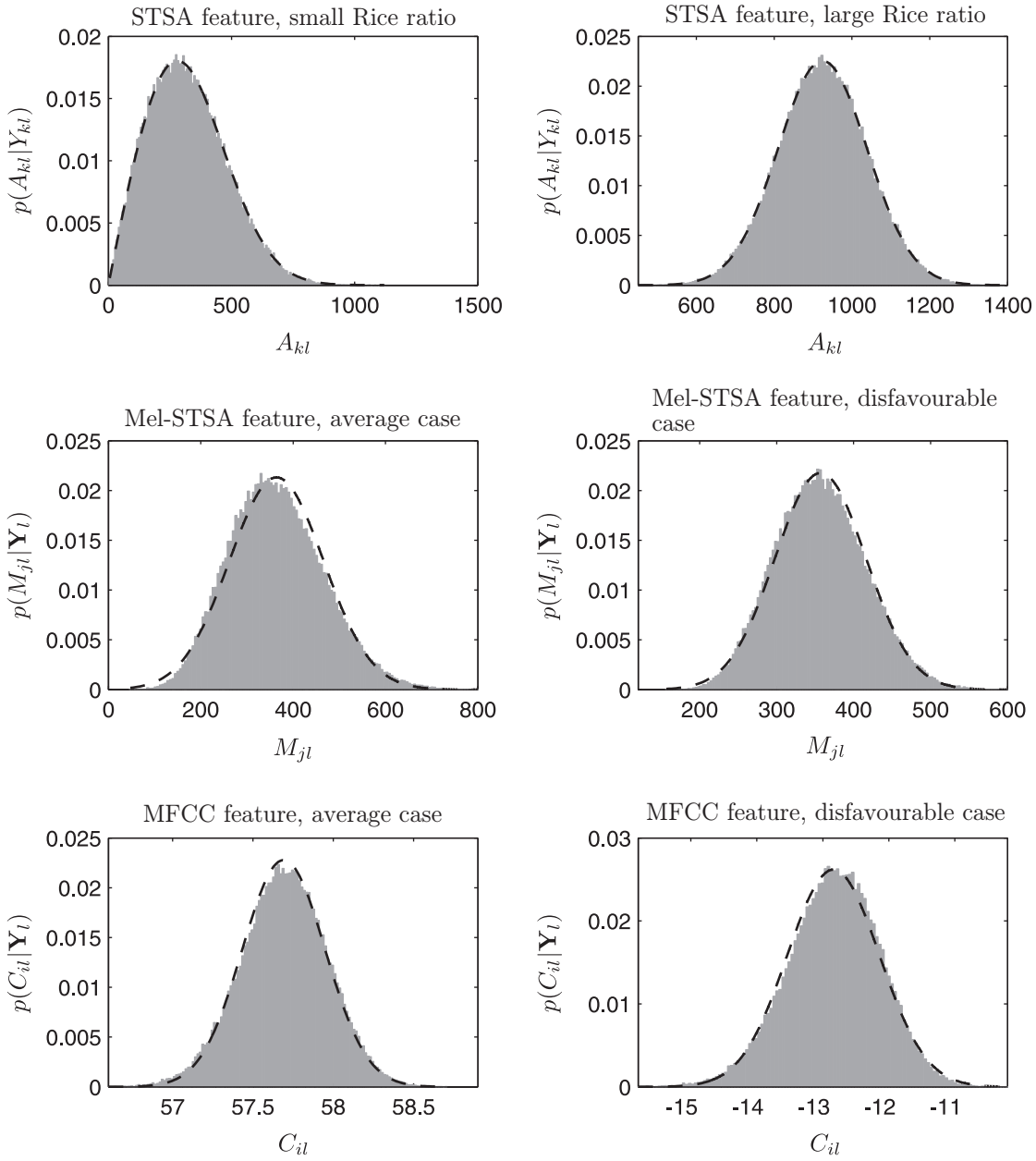

Figure 7.3: **From top to bottom**: Empirically obtained distribution (shaded) and hypothesized distribution (dashed line) of uncertainty of the STSA, Mel-STSA and MFCC features. An example representative of the average case is presented on the left, whereas an example representative of the most unfavorable cases is represented on the right. In the case of the STSA features an example of a small Rice ratio is displayed on the left and an example of a large Rice ratio on the right.

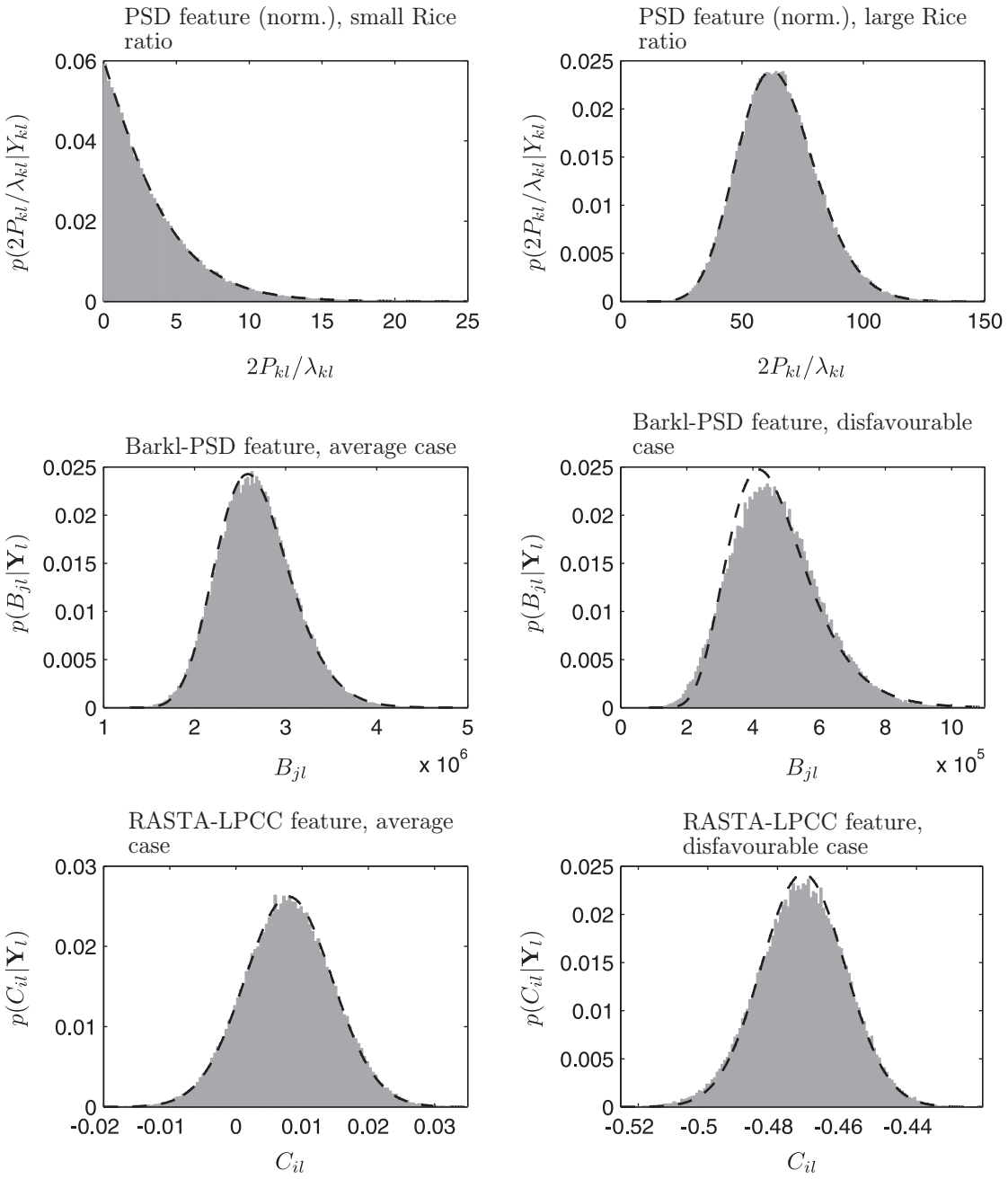

Figure 7.4: **From top to bottom**: Empirically obtained distribution (shaded) and hypothesized distribution (dashed line) of uncertainty of the PSD, Bark-PSD and RASTA-LPCC features. An example representative of the average case is presented on the left, whereas an example representative of the most unfavorable cases is represented on the right. In the case of the PSD features an example of a small Rice ratio is displayed on the left and an example of a large Rice ratio on the right. PSD features are also normalized.

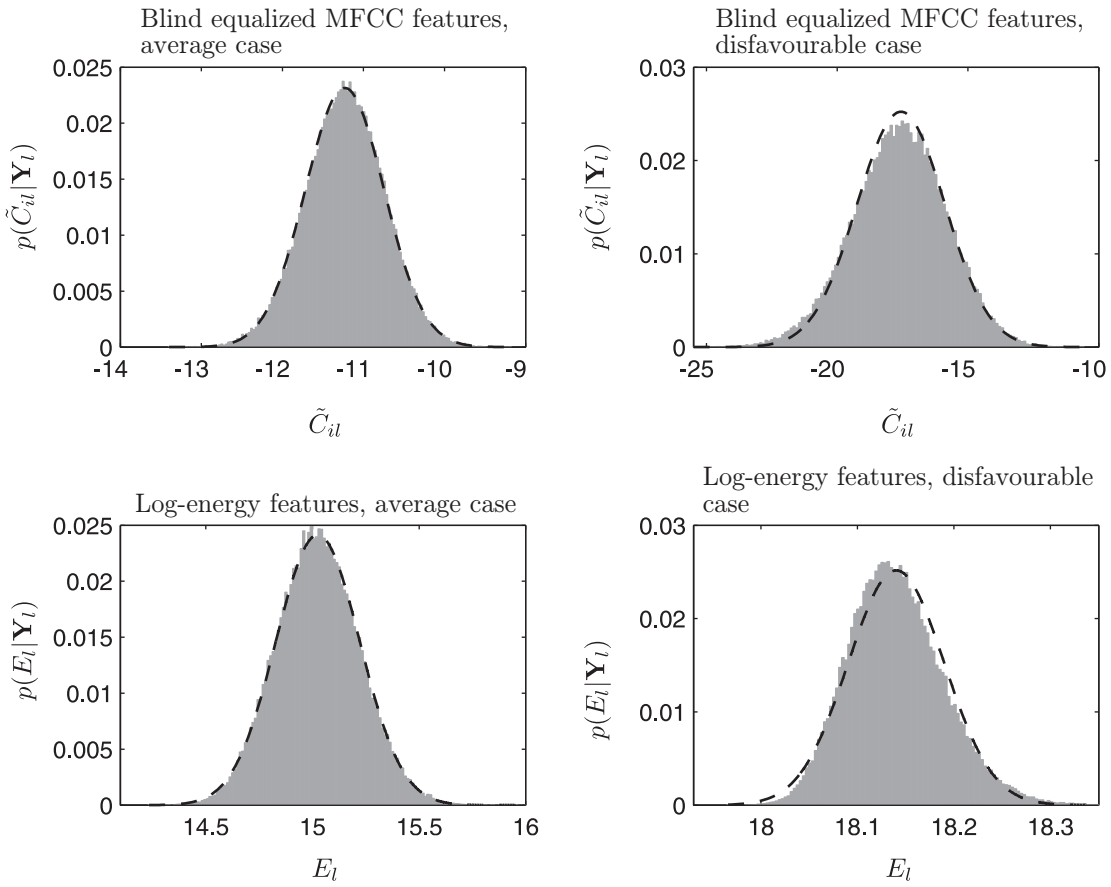

Figure 7.5: **From top to bottom**: Empirically obtained distribution (shaded) and hypothesized distribution (dashed line) of uncertainty of the blind equalized cepstra (BE) and Log-energy (LogE) features. An example representative of the average case is presented on the left, whereas an example representative of the most unfavorable cases is represented on the right.

7.1.2 Accuracy of the Developed Uncertainty Propagation Algorithms

The Montecarlo simulation was also used to test the efficiency of the proposed uncertainty propagation algorithms. Both in the case of the MFCC and RASTA-PLP uncertainty propagation algorithms, the most relevant approximation corresponds to the propagation of uncertainty through the logarithm transformation (Log-Mel-STSA and Log-Bark-PSD transformations). This is performed either with the unscented transform (see Section 6.2.3) or the log-Normal assumption (see Section 6.3.3) respectively. The accuracy of estimation in this transformation was therefore analyzed along with the accuracy obtained for the final features of the MFCCs (DCT), RASTA-LPCCs (LPCC) and the ETSI AFE (LogE and BE).
Three alternatives were compared: Propagated mean and covariance computed from the Montecarlo simulation, propagated mean and covariance computed using the uncertainty propagation algorithms proposed in Chapter 6 (labeled F) and the same algorithms using only diagonal covariances (see Section 6.5.1) (labeled D).

7.1.2.1 Results

A representative example of the test results is displayed in Figures 7.6 and 7.7. In general, and regardless of the SNR, the approximation error for the estimated means was very low. In most of the scenarios, both Montecarlo and the implemented algorithm estimations were indistinguishable. The accuracy of the mean estimation was also not affected by the use of full (F) or diagonal covariances (D). The error in covariance estimation when using full covariance in all scenarios was also rather low. For this reason and for clarity issues, only the covariance of the feature with the highest absolute estimation error is displayed. This error was computed by comparing Montecarlo approximation and full covariance uncertainty propagation.
The use of the unscented transform for computation of the Log-Mel-STSA variance (Figure 7.6, top) led to a slight overestimation of the variance. This overestimation is only noticeable in some features and its effect on the final MFCC features (Figure 7.6, middle) is reduced due to the averaging effect of the DCT transformation (see Section 6.2.4). The use of diagonal covariance or full covariance after the Mel-filterbank produces a slight dissimilitude in the propagated Log-Mel-STSA variance due to implementation issues of the unscented transform. In the case of the final MFCC features (Figure 7.6, middle), the use of diagonal covariance causes an underestimation of the MFCC variance, although it remains approximately proportional to the full covariance case.
The use of the log-Normal approximation for the computation of the Log-Bark-PSD variance (Figure 7.6, bottom) leads to an underestimation of the variance with a slightly higher error than in the unscented transform case. This error does however affect the final LPCC features much less than in the MFCC case (Figure 7.7, top). The reduced influence of this error is probably again due to the averaging performed in the multiple posterior transformations. The opposite phenomenon can be observed with the variance estimation using only diagonal covariance, which leads to a strong overestimation of the variances in some of the features. This does not however affect much the efficiency of

the algorithm as it will be seen in the next section.
The use of the log-Normal approximation for the computation of the LogE features (Figure 7.7, middle) exhibits a lower error than in the case of the Log-Bark-PSD features (Figure 7.6, bottom). This is likely due to the strong influence of the summation operation in Equation 3.28. This operation has an averaging effect which seems to favor the convergence to the log-Normal distribution. The estimation error when using diagonal covariances (D) after the IDFT, is however higher than in the filterbank case. This can be explained by the fact the IDFT transformation corresponds to a full matrix whereas in the case of the filterbank matrix only neighboring elements are correlated (see Figure 3.3). Ignoring the correlation induced by the IDFT leads therefore to a higher variance estimation error. With respect to the blind-equalized cepstra (Figure 7.7, bottom), the result is very similar to the conventional MFCC case with the full covariance propagation displaying a high accuracy and the diagonal case leading to underestimation but maintaining approximate proportionality.

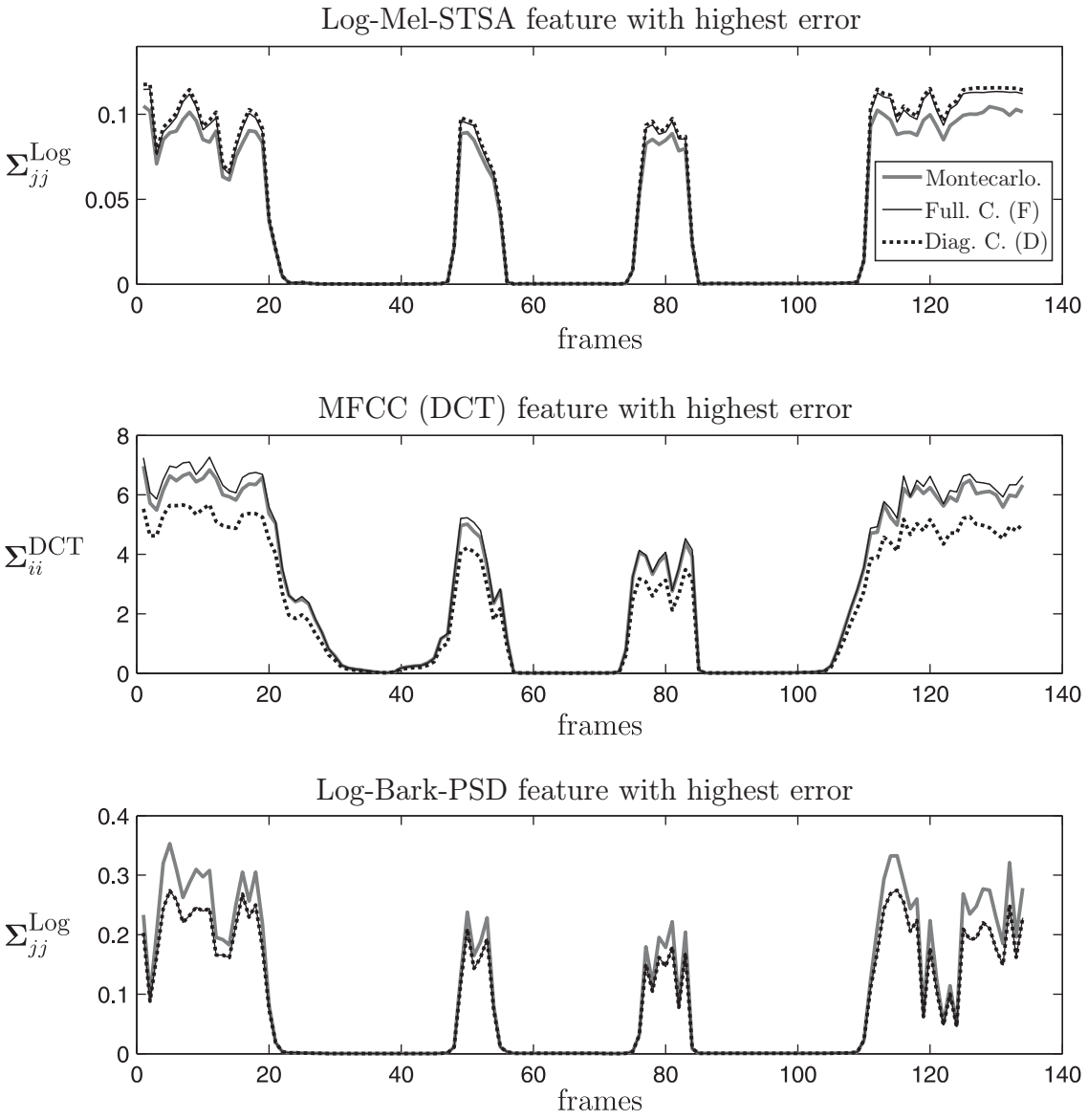

Figure 7.6: Comparison between propagated variance computed using a Montecarlo simulation (thick gray) with the proposed uncertainty propagation algorithms using diagonal (dotted black) and full covariance (solid black). Only the feature with the highest absolute variance estimation error when comparing Montecarlo and (F) option is displayed. **From top to bottom:** Log-Mel-STSA, MFCC (DCT), Log-Bark-PSD. Note that the accuracy of the propagated Log-Bark-PSD variance is not affected by the use of full or diagonal covariances. Both solutions consequently overlap.

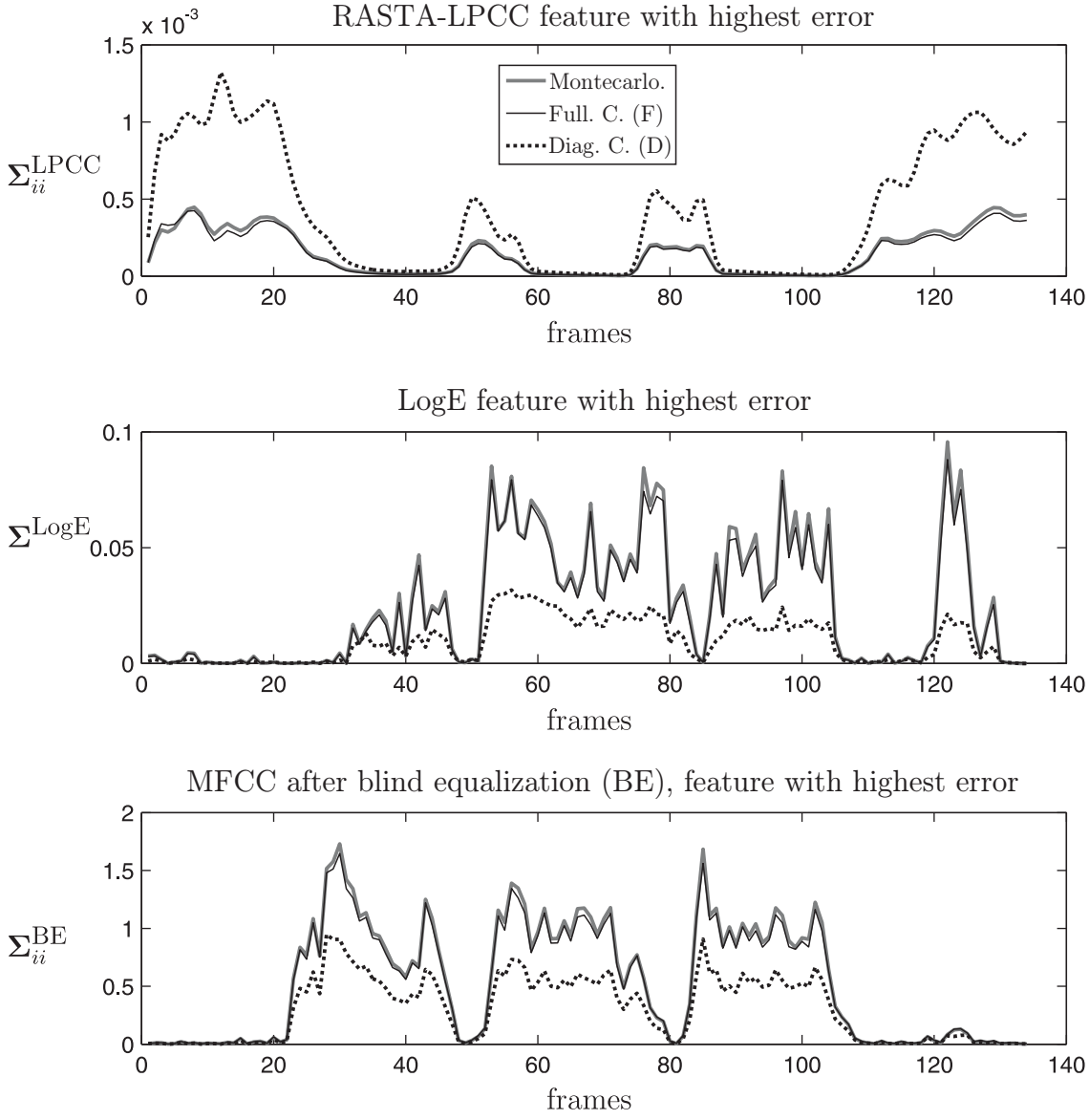

Figure 7.7: Comparison between propagated variance computed using a Montecarlo simulation (thick gray) with the proposed uncertainty propagation algorithms using diagonal (dotted black) and full covariance (solid black). Only the feature with the highest absolute variance estimation error when comparing Montecarlo and (F) option is displayed. **From top to bottom:** RASTA-LPCC, LogE, BE (MFCC).

7.2 Tests with the AURORA5 Database and the ETSI ES 202 050 Standard for Robust Speech Recognition

7.2.1 AURORA5 Noisy Speech Database

In order to test the developed uncertainty propagation algorithms in a realistic ASR environment, the test framework of the AURORA5 noisy speech database [Hirsch 2007] was used. The AURORA tasks comprise noisy versions of well known clean speech databases like the TI-DIGITS [TIDigits 1993] or Wall-Street-Journal [Hirsch 2002]. They aim to provide a normalized framework for the test of robust automatic speech recognition systems [Hirsch 2000]. The AURORA5 release is particularly designed to work with the TI-DIGITS database and the advanced front-end (AFE) specified in the ETSI standard ES 202 050 and introduced in Section 6.4. It also contains train and test batch scripts for its use with the hidden Markov model toolkit (HTK), a toolbox of the Cambridge University for the training and testing of HMM based ASR systems [Young 2006].
The speech recognition task comprises a fully connected language model containing the words 'one', 'two', 'three', 'four', 'five', 'six', 'seven', 'eight', 'nine', 'zero' and 'oh'. The original TI-DIGITS database contains recordings of 225 different speakers, for each of which 77 utterances were recorded. The AURORA5 downsamples the original 20KHz files of TI-DIGITS to 8KHz. The model template provided by the AURORA5 database is a set of 11 word models with a left-right HMM structure of 16 states plus a 5 state silence model. The number of mixtures per state is 4. The training procedure follows the standard 'flat start' training scheme as specified in the HTK manual [Young 2006]. A high number of re-estimations, 10, is however used after each mixture splitting. Since the AURORA5 contains multiple tests for different purposes, only some of the tests were performed. The ASR tests carried out are fully described in [Hirsch 2007, sec. 5.3] and were chosen due to their generality. They correspond to training with clean data and testing with noisy speech with five different levels of SNR (0, 5, 10, 15, ∞) and three of reverberance (none, office, living room)[2]. The HTK training/test scripts available in the AURORA5 database were translated from Linux to MATLAB code for use on a Windows machine. In order to perform recognition with uncertain features, the HTK executable HVite was modified to perform uncertainty decoding and modified imputation as given by Equations 3.40 and 3.44. Only the use of diagonal covariances for recognition was implemented.

7.2.2 Implementation of Uncertainty Propagation for the ETSI Advanced Front-End

To test uncertainty propagation on the AURORA5 task, the complete advanced front-end algorithm was implemented in MATLAB using the standards manual [ETSI 2007] and the source code for version 1.1.2 (available online) as reference [Mandelartz 2009].

[2]Note that rather than using segmental SNR as in Equation 7.2, or another type of scaling factor, AURORA5 uses the software in [Hirsch 2005], to simulate different environment acoustics.

This standard was later modified to allow piecewise uncertainty propagation as described in Chapter 6.
Due to the complexity of the speech enhancement steps mentioned before, measures of uncertainty based on the MMSE estimation error (see Section 5.3) are not available. The variance of the uncertainty model was computed based on empirical methods (see Section 5.2 for details). The estimator is given in Equation 5.18.
Although an accurate approximation for the propagation through the log-energy feature extraction was proposed in Section 3.3.2, this was not included in the final version of the testing framework due to various reasons. Despite the accuracy of the propagation approach, the results were slightly worse when considering the energy features as uncertain. Its use also posed additional difficulties when determining the uncertainty of estimation using Equation 5.18 since the log-energy is computed from the time domain signal without pre-emphasis (see Section 5.2). Finally, the consideration of the log-energy as deterministic was coherent with the same assumption being made for the propagation through the blind equalization transformation (see Section 6.4.3). For these reasons the log-energy was considered as deterministic in the ASR tests.

7.2.3 Speech Recognition Results

The AURORA5 test results are displayed in Table 7.2. The first column specifies the SNR, the second corresponds to the baseline results obtained with the MATLAB implementation of the ETSI standard. The next four columns correspond to the use of uncertainty propagation in addition to the ETSI standard, combined either with modified imputation (MI) or uncertainty decoding (UD) and using either full covariance (F) or diagonal covariance (D). From the obtained results, it is clear that both modified imputation and uncertainty decoding, combined with uncertainty propagation, reduce the word error rate (WER) (see Section 3.5) of the AFE in all scenarios with the only exception of the clean non-reverberant case, which experiences a very small increase in the error rate. The relative reduction of the error rates for the non-reverberant environments for which the standard was optimized, is small, around an 8% reduction on average for the best combination (MI+D). This can be expected since the standard is a highly optimized algorithm with excellent speech enhancement characteristics. Its performance in a controlled environment is therefore difficult to improve. This is not the case for the reverberant scenarios. Error rates in the reverberant environments without presence of noise are reduced by as much as 34% and 17% for the low and high reverberance environments respectively. Performance in reverberant environments under the presence of noise is also improved by employing uncertainty propagation. Maximal reductions in WER of $10\% - 30\%$ and $8\% - 20\%$ are attained with option (MI+F) for the two reverberant environments respectively.
Regarding the use of modified imputation or uncertainty decoding, the former method showed a superior performance in all the tests with the only exception of the clean speech or very high SNR cases. Modified imputation often doubled the relative reduction in WER attained with uncertainty decoding in reverberant environments.
The use of diagonal or full covariances for propagation had a varying but limited effect. For reverberant scenarios the use of full covariances had in general a positive effect in

Table 7.2: Word error rates [%] for the advanced front-end baseline (AFE), its combination with uncertainty propagation and modified imputation (MI) or uncertainty decoding (UD). Use of diagonal or full covariance by propagation marked as (D) or (F) respectively. The best results are displayed in bold.

Non Reverberant					
SNR	AFE	+MI+D	+UD+D	+MI+F	+UD+F
∞	0.51	0.65	0.54	0.70	0.54
15	2.85	2.65	2.66	2.74	**2.62**
10	5.97	**5.52**	5.65	5.62	5.65
05	14.60	**13.22**	13.56	13.50	13.49
00	34.32	**31.55**	32.83	31.80	32.64
Hands Free Office					
SNR	AFE	+MI+D	+UD+D	+MI+F	+UD+F
∞	5.85	3.94	4.86	**3.86**	4.72
15	11.06	7.80	9.38	**7.46**	9.10
10	17.54	13.45	15.47	**13.01**	15.04
05	30.29	25.10	27.90	**24.47**	27.40
00	51.36	45.89	48.69	**45.15**	48.33
Hands Free Living Room					
SNR	AFE	+MI+D	+UD+D	+MI+F	+UD+F
∞	14.27	**11.82**	13.22	11.97	12.94
15	20.54	16.39	18.72	**16.19**	18.41
10	28.60	23.56	26.40	**23.12**	25.97
05	42.73	37.16	40.34	**36.70**	39.88
00	62.32	57.40	60.17	**56.78**	59.84

the recognition rates whereas its use in non-reverberant environments led to slightly poorer recognition rates. However, the difference between the use of full and diagonal covariances remained small in all scenarios.

7.3 Tests with the AURORA5 Database and AFE-RASTA Features

In addition to the AURORA5 tests employing the complete standard, tests were carried out using the RASTA-LPCC feature extraction introduced in Section 3.3.2 and the corresponding uncertainty propagation algorithm proposed in Section 6.3. For this purpose the standard introduced in Section 3.4 was modified as follows. The MFCC feature extraction performed in step 3 was replaced by the RASTA-LPCC feature extraction explained in Section 3.3.2. Since no codebook was available for the RASTA-PLP cepstras, the blind equalization in step 4 was replaced by cepstral mean subtraction (described in Sections 3.3.6 and 6.5.2). The transmission aspects in step 5 were ignored. The server side computations in step 6 were also replaced by conventional first order delta

and acceleration coefficients. The RASTA-PLP cepstra implementation was based in the one of the RASTAMAT toolbox [Ellis 2005]. Both diagonal (D) and full covariance (F) uncertainty propagation were tested.

7.3.1 Speech Recognition Results

The results of the tests, displayed in Table 7.3, show a reduction of baseline WER similar to the one obtained with the conventional advanced front-end. However the RASTA-LPCC features baseline exhibits a much lower robustness to reverberance than the original MFCC implementation. It has to be taken into account that the original implementation included both a blind equalization scheme and a server side frame dropping system which increased their robustness. The speech enhancement of the ETSI standard was also most surely optimized for the original MFCC features.
As in the MFCC case, improvements in the non-reverberant noisy environments are very small, lower than a 4% in average for the best combination (MI+F). WER increases slightly in the clean speech case for the (MI) combinations and experiments a small

Table 7.3: Word error rates [%] for the advanced front-end baseline with RASTA-LPCC features (RASTA), its combination with uncertainty propagation and modified imputation (MI) or uncertainty decoding (UD). Use of diagonal or full covariance by propagation marked as (D) or (F) respectively. The best results improving the baseline are displayed in bold.

Non Reverberant					
SNR	RASTA	+MI+D	+UD+D	+MI+F	+UD+F
∞	0.80	0.84	**0.76**	0.82	0.77
15	4.88	4.88	4.89	**4.81**	4.93
10	9.89	9.83	9.88	**9.53**	9.80
05	21.76	21.31	21.46	**20.54**	21.38
00	45.30	44.00	44.77	**43.15**	44.75
Hands Free Office					
SNR	RASTA	+MI+D	+UD+D	+MI+F	+UD+F
∞	22.83	**17.52**	20.54	18.73	21.46
15	31.03	**25.14**	28.66	26.24	29.39
10	38.91	**33.44**	36.88	34.25	37.32
05	51.75	**47.09**	50.10	47.5	50.43
00	69.05	65.98	67.95	**65.90**	68.15
Hands Free Living Room					
SNR	RASTA	+MI+D	+UD+D	+MI+F	+UD+F
∞	38.75	**33.90**	36.94	35.01	37.50
15	46.76	**41.36**	44.87	42.22	45.45
10	54.06	**49.04**	52.37	49.89	52.78
05	64.68	**60.68**	63.39	61.20	63.73
00	77.57	**74.82**	76.64	74.98	76.76

decrease for the (UD) combinations. The reduction in WER attained in the reverberant environments is higher than in the non reverberant case, with a $4\%-23\%$ and a $4\%-13\%$ reduction for the best combination (MI+D). The relative improvement decreases for lower SNRs.

7.4 Tests using GMM Priors for Speech and Noise

7.4.1 Speech with Artificially Added Noise

In addition to the AURORA5 tests specified in the standard, a set of tests were carried out by artificially adding noise to the AURORA5 clean test data using the procedure introduced in Section 7.1. These additional tests were needed to test implementations based on conventional versions of the MMSE estimators, like the ones developed in Chapters 4 and 5. Methods like MMSE-STSA or MMSE-LSA performed poorly on the noisy files of the AURORA5 database resulting in high WERs for the available range of SNRs. By using artificially added noises, the SNR could be selected to obtain sufficiently high baseline results. This procedure also provided a better insight into the underlying noise process useful for research purposes.
The AURORA5 model templates, training and test routines remained the same in all the tests, although the amount of data used for testing varied.

7.4.2 Implemented Speech Enhancement Baseline

The speech enhancement scheme implemented corresponds to the one presented in Chapter 2. Estimation of SNR (see Section 2.2) was performed using recursive smoothing driven by a speech presence probability estimation as given by Equations 2.54 and 2.55. The speech probability was determined using minimum statistics as specified in the IMCRA noise variance estimation method [Cohen 2001]. This was combined with conventional decision directed a priori SNR estimation given by Equation 2.59 with $\alpha = 0.92$. Clean speech estimation was performed using MMSE-STSA and MMSE-LSA estimators given by Equations 2.41 and 2.47. Since those methods use Gaussian priors for speech and noise, they were labeled (GS+GN). The two types of noise added were Gaussian noise, and instationary wind noise recorded on a vehicle on a highway. The feature extraction used was the Mel-cepstra described in Section 3.3.1 with an additional preemphasis (see Section 3.3.4) of 0.97, delta and acceleration features (see Section 3.3.5). Unlike in the original AURORA5 configuration, the speech enhancement method was not used during training. Training and testing was carried out using the complete clean sets (around 8600 files each).

7.4.3 Implemented Supergaussian Estimators

Two versions of the developed RiMM-MMSE-STSA estimator given by Equation 4.42 and RiMM-MMSE-LSA estimator given by Equation 4.45 were tested. Firstly a Gaussian speech prior was combined with the GMM noise prior computed with the method introduced in Section 4.4. A number of $J = 5$ mixtures was used. This was labeled as

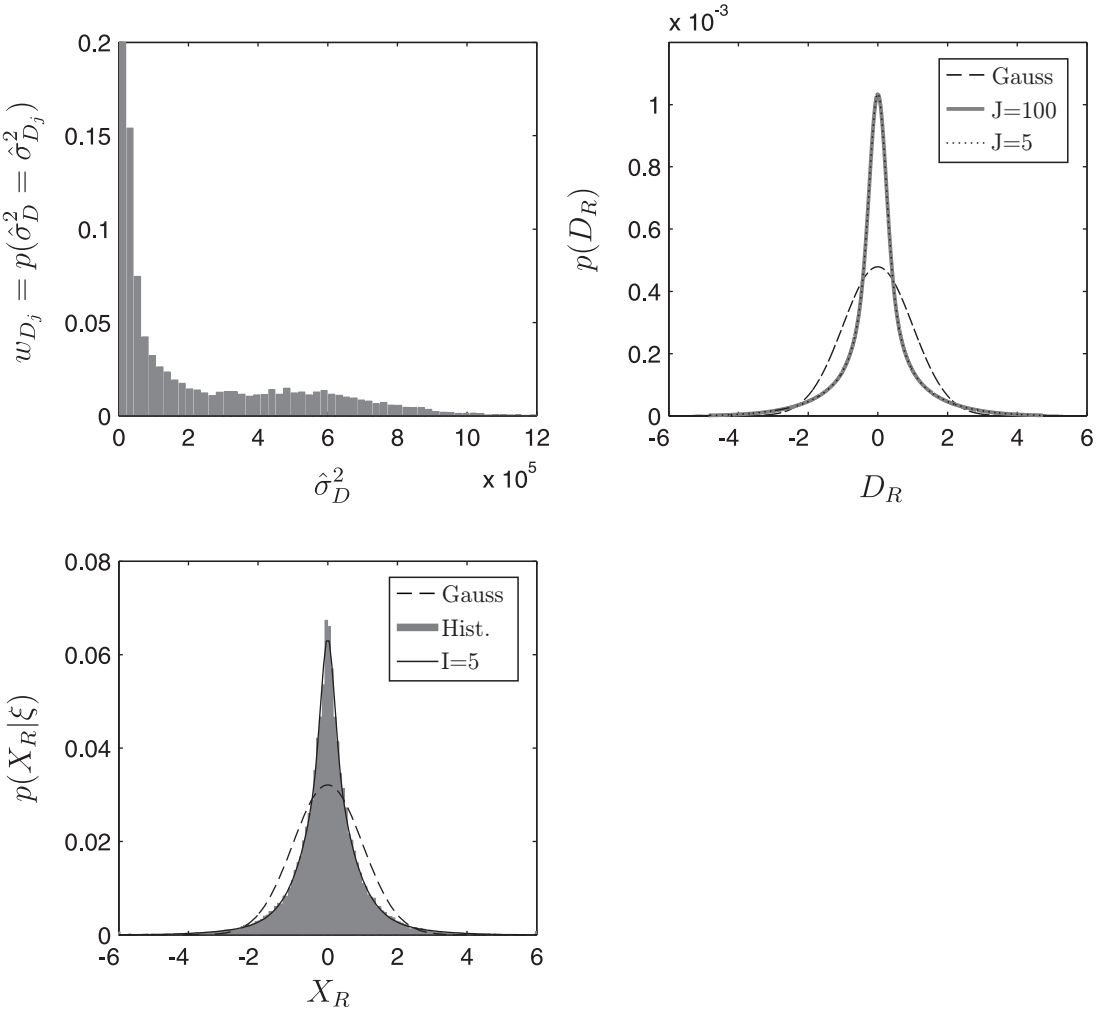

Figure 7.8: **Top, left:** Empirically obtained distribution of the estimated variance of the real Fourier component of noise $\hat{\sigma}_D^2$ (shaded). **Top, right:** Gaussian prior of the real component of noise D_R (dashed). $J = 100$ mixtures GMM noise prior of D_R corresponding to the histogram on the left (solid gray). $J = 5$ mixtures GMM noise prior trained with Montecarlo simulation (black dotted). **Bottom, left:** Empirically obtained distribution of the real component of the speech Fourier coefficient X_R, samples were drawn for a range of ξ of 19dB-21dB (shaded). Gaussian prior for X_R (dashed). $I = 5$ mixtures GMM speech prior trained with the obtained samples (solid).

(GS+SN). Secondly the same GMM noise prior was combined with a GMM speech prior of $I = 5$ mixtures. This is equivalent to the Rayleigh mixture model amplitude prior used in [Erkelens 2007]. This experiment was consequently labeled (SS+SN).
The parameters of the prior of each noise Fourier coefficient $p(D)$ were determined from the parameters of its real component $p(D_R)$. This is displayed at the top-right in Figure 7.8. As as explained in Section 4.4, the distribution of the estimated variance

Table 7.4: Word error rates [%] for the implemented baseline when no speech enhacement is used (Noisy). Gaussian prior baseline (GS+GN). Gaussian prior for speech and supergaussian noise prior for noise (GS+SN). Supergaussian priors for noise and speech (SS+SN). Results that improve the Gaussian prior baseline (GS+GN) are displayed in bold.

Gaussian-Noise							
		MMSE-STSA			MMSE-LSA		
SNR	Noisy	GS+GN	GS+SN	SS+SN	GS+GN	GS+SN	SS+SN
∞	1.39	1.41	1.43	1.49	1.53	**1.50**	1.86
10	7.71	3.13	3.33	4.01	3.30	**2.86**	4.24
05	16.59	5.71	6.44	8.18	5.79	**4.94**	7.58
00	29.26	12.53	14.29	17.77	12.95	**10.59**	15.56
Wind-Noise							
		MMSE-STSA			MMSE-LSA		
SNR	Noisy	GS+GN	GS+SN	SS+SN	GS+GN	GS+SN	SS+SN
∞	1.39	1.41	1.43	1.49	1.53	**1.50**	1.86
05	2.00	2.02	**1.75**	2.24	2.32	**1.89**	2.66
00	3.93	3.62	**2.69**	**3.61**	3.84	**2.93**	**3.77**
-05	9.89	8.98	**6.77**	**8.42**	8.87	**6.85**	**7.28**

$p(\hat{\sigma}_D^2)$ was approximated by a 100 bin histogram. Samples of $\hat{\sigma}_D^2$ were obtained from the IMCRA estimator [Cohen 2003] in speech enhancement experiments using the baseline configuration. For this purpose, 20 clean files, to which noise samples had been added, were used. The noise samples included Gaussian[3] and instationary train noise from the JEIDA database [Itahashi 1990] (not used in the test set), added at an SNR of 0dB. Both MMSE-STSA and MMSE-LSA estimators were used. Once the 100 bin histogram was computed (see Figure 7.8, top-right), this corresponded to a $J = 100$ mixture prior as given by Equation 4.54. To reduce the number of mixtures needed, Montecarlo simulation was used. A $J = 5$ mixture GMM was trained by drawing 10^6 samples from the $J = 100$ mixture GMM and using the EM algorithm [Press 2007, Sec. 16.1]. This resulted in an almost identical prior. Both priors overlap completely in Figure 7.8.
The speech prior was determined following the criterion used in [Erkelens 2007] and [Lotter 2005] but with a slightly different procedure. A set of 80 clean speech files[4] was processed with the baseline configuration. As in the original paper, samples of speech were restricted to those obtained for an a priori SNR range of 19dB-21dB. With the collected samples, an $I = 5$ mixtures Gaussian mixture model was trained using the real component of each speech Fourier coefficient and the EM algorithm. The resulting prior is displayed at the bottom of Figure 7.8.

[3]Note that a noise prior obtained from noise variance estimations of Gaussian noise is not Gaussian. This apparent contradiction is due to the imperfection of the variance estimation.
[4]More files than in the noise prior case are needed since only some samples satisfy the condition in Equation 4.4.

7.4.4 Speech Recognition Results

Results displayed in Table 7.4 show a clearly different behavior for RiMM-MMSE-STSA and RiMM-MMSE-LSA estimators when employing only the supergaussian noise prior (GS+SN). In the case of RiMM-MMSE-STSA, error rates increase by 11% on average in the Gaussian noise scenarios whereas an error reduction of 13%-26% is obtained for the instationary noise. RiMM-MMSE-LSA however shows a coherent improvement of 13%-23% for both scenarios, including even a small improvement for the clean speech case. In both cases, the higher relative reductions in WER are obtained at the lower SNRs.
When combined with the supergaussian speech prior both RiMM-MMSE-STSA and RiMM-MMSE-LSA exhibit a clearly poorer performance. The WER increases in almost all scenarios with respect to the conventional MMSE estimator. The only scenario, in which results improve the baseline are for instationary noises and low SNRs. Improvements in these scenarios reach a maximum of 17%.

7.5 Tests using MMSE Posterior Propagation

7.5.1 Implemented Speech Enhancement Baseline

The configuration used for these tests is similar to the one used in the previous section. In accordance with the algorithms developed in Chapter 5, baseline results were computed for Wiener, MMSE-STSA and MMSE-LSA estimators as given by Equations 2.30, 2.41 and 2.47. For this scenario, the version of the algorithms incorporating speech presence uncertainty was used with $G_{\min} = 0.1$ (see Section 2.1.6). Conventional Bayesian error was used as the measure of uncertainty regardless of speech presence probability. Speech enhancement[5] was also employed during training, which had the effect of adapting the trained models to the artifacts caused by the speech enhancement methods when applied to clean speech. The use of this form of static adaptation was desirable in this case for two main reasons. On the one hand, uncertainty propagation can be seen as a method for dynamic compensation of HMMs. Adapting the model to the clean speech artifacts ensures that the effect of uncertainty propagation goes beyond this easily obtainable form of compensation. On the other hand at high SNRs, the Wiener filter combined with IMCRA attained a performance similar to that of the MMSE-STSA and MMSE-LSA methods[6]. However, the use of this training scheme had the effect of increasing the general baseline WER for all the methods.
Tests were carried out using 10 speakers, half male, half female, with a total of 770 files. This was found to be representative of the total test-set and allowed a considerable reduction in the test execution times. The other configuration aspects of the baseline were not changed.

[5]Without consideration of speech presence uncertainty.
[6]Since IMCRA needs of the computation of ξ it is affected by the clean speech estimation method used. IMCRA was originally designed to work with the MMSE-LSA estimator.

Table 7.5: Word error rates [%] for Wiener, MMSE-STSA (STSA), MMSE-LSA (LSA) estimator baselines and its combination with uncertainty propagation and modified imputation (MI) as well as uncertainty decoding (UD). Best results are displayed in bold.

Gaussian-Noise									
SNR	Wiener	+MI	+UD	STSA	+MI	+UD	LSA	+MI	+UD
∞	1.39	1.35	**1.31**	1.39	**1.35**	1.39	1.35	**1.31**	1.35
15	2.02	**1.78**	1.86	1.86	**1.67**	1.74	1.90	**1.70**	**1.70**
10	3.89	**2.89**	2.97	3.69	**2.66**	2.78	3.89	**2.89**	2.97
05	10.31	**7.73**	8.37	10.07	**7.06**	8.01	10.03	**7.61**	8.33
00	19.43	**15.42**	16.65	19.15	**15.03**	16.42	19.19	**15.86**	17.17
Crowded-Street Noise									
SNR	Wiener	+MI	+UD	STSA	+MI	+UD	LSA	+MI	+UD
∞	1.39	1.35	**1.31**	1.39	**1.35**	1.39	1.35	**1.31**	1.35
10	2.74	**2.06**	2.14	2.38	**2.14**	**2.14**	2.38	**2.06**	2.10
5	5.87	**3.85**	4.12	4.84	**3.61**	3.81	5.19	**3.69**	3.89
0	11.93	**9.56**	9.83	11.34	**8.60**	9.16	11.10	**8.96**	9.28
-5	20.98	**18.00**	18.32	21.29	**17.72**	18.16	20.94	**17.76**	18.32

7.5.2 Implemented Posterior Propagation Methods

The complex Gaussian model was used for describing the uncertain STFT estimation obtained from the Wiener, MMSE-STSA and MMSE-LSA estimators as given by Equations 5.29, 5.40 and 5.41. Since the Mel-cepstral feature extraction, introduced in Section 3.3.1, was used, the corresponding uncertainty propagation scheme introduced in Section 6.2 was applied. Diagonal covariance was used for propagation. The recognition under observation uncertainty was performed using both modified imputation, labeled MI in the results, and uncertainty decoding, labeled UD. Both techniques were introduced in Section 3.6.

7.5.3 Speech Recognition Results

As shown in Table 7.5, the use of uncertainty propagation improves recognition results in all noisy scenarios regardless of SNR or uncertainty of observation technique used. The results on clean speech are also either improved or equaled. The use of modified imputation reduces error rates in the noisy scenarios by $10\% - 34\%$ with an average of 21%. A slight improvement is also obtained for clean speech. As in the case of the ETSI AFE tests, modified imputation tops uncertainty decoding performance with the only exception of the clean speech cases. WER reduction when using uncertainty decoding ranges from 7% to 30% with an average of 17%. The positive effect of both techniques peaks at 10dB SNR.

8 Discussion and Conclusions

8.1 Discussion

The uncertainty propagation technique presented in this thesis can be divided into three main processes or steps, estimation of uncertainty (Chapter 5), propagation (Chapter 6) and recognition (Section 3.6), each of which is not ideally solved. When trying to analyze the efficiency of the approach, errors in one of the processes might obfuscate the performance issues of the other two.
If we assume the complex Gaussian uncertainty model in the Fourier domain as a valid premise, the accuracy of uncertainty propagation can be assessed independently of the estimation and recognition steps by using Montecarlo simulations. This has been done in the experimental setup in Section 7.1 and is commented on this chapter in Section 8.1.1.
Regarding uncertainty estimation and recognition under observation uncertainty, a problem arises. In theory, an ideal measure of uncertainty could be achieved if the clean signal X is available as

$$\lambda_{kl} = |X_{kl} - \hat{X}_{kl}|^2. \tag{8.1}$$

However, although the propagation of this ideal measure of uncertainty leads to a reduction of WER much superior to the one obtained with realistic variance estimators, it does not provide the optimal results. Better results can be obtained by using transformations of this ideal variance, like for example lowering or ignoring variances generated when speech is not attenuated ($\hat{X}_{kl} > X_{kl}$). This phenomenon seems to be caused by the particular way in which modified imputation and uncertainty decoding incorporate uncertainty to improve the recognition process, which is imperfect and was discussed in Section 8.1.2.
As a consequence of this, no ideal measure of uncertainty is available to judge the optimality of the empirically obtained measures of uncertainty (see Section 5.2). The efficiency of these estimations in improving ASR, demonstrated in the experimental setup in Sections 7.2.3 and 7.3.1, must therefore be circumscribed to the context of uncertainty propagation combined with modified imputation and uncertainty decoding.
This is not the case with the measures of uncertainty based on the posterior distribution of the speech enhancement model (see Section 5.3), which are discussed particularly in Section 8.1.5.

8.1.1 Piecewise Uncertainty Propagation

Although the piecewise uncertainty propagation approach, proposed in Chapter 6, has only been applied to MFCC, RASTA-PLP and ETSI AFE features, it can be used to

yield similar propagation approaches for other feature extractions. The piecewise approach can be seen as set of 'recipes' that can be combined or extended to achieve uncertainty propagation for multiple transformations.
Propagation through feature extractions using filterbanks (i.e Mel, Bark), combined with given powers of the STSA (i.e. PSD), can be solved using the closed form solution for the Rice moments (Equation 6.4) and the linearity of the filterbanks. The provided closed form solution allows the computation of all the needed cumulants of the uncertain filterbank features[1].
After the filterbank transformation, the dimensionality of the features has been reduced by an order of magnitude, which allows the use of pseudo-Montecarlo techniques. As long as the distribution of the uncertain filterbank features is not very skewed (i.e. STSA features), the unscented transform has been demonstrated to provide a good approximation of propagated mean and covariance with a low consumption of resources (see Section 6.2.3.1). This method is also not limited by the complexity of the posterior transformations between features. One exception to this case are transformations that combine features from different frames. In such a case the dimensionality of the features to be propagated is multiplied by the number of frames, thus rendering the unscented transform ineffective again. For the particular case IIR filters (i.e. RASTA filter), closed form solutions similar to the one presented in Section 6.3.4 can be obtained.
If the uncertain features after the filterbank are too skewed to be propagated with the unscented transform (i.e. PSD features), the log-Normal assumption might provide a good representation of the distribution. In the particular case of the PSD features, the log-Normal assumption has been demonstrated to be a reasonable approximation (see Section 6.3.3). The condition of log-Normality is not altered by multiplication by a scalar, or exponentiation, and the resulting statistical parameters are easy to compute. This simplifies the further propagation of uncertain filterbank features and would allow, for example, the computation of the conventional PLP features.
Finally, inverting the complex Gaussian uncertain Fourier spectrum back into the time domain yields a Gaussian distributed variable (see Section 6.4.1). Its log-energy has also demonstrated to be well approximated by a Gaussian distribution (see Section 6.4.2). Both closed form solutions might be employed with feature extractions starting in the time domain.

8.1.2 Limitations of Modified Imputation and Uncertainty Decoding

As noted by Liao and Gales [Liao 2008], both uncertainty decoding and modified imputation suffer from a limitation that might explain their behavior when using ideally computed uncertainties. Both methods implicitly assume that the distribution of feature uncertainty is independent of the mixture being evaluated and thus of the word model. In case of input frames with a high uncertainty, all mixture probability computations are strongly and equally modified. This masks the individual speech characteristics of each model and 'floods' the HMM ASR system with uncertainty. Without the help of other discriminative information, like word transition probabilities, such behaviour can

[1] A closed form solution for the cumulants of the generalized non-central Chi model of uncertainty (see Section 5.1.2) can equivalently be obtained.

lead to a poor recognition performance even if the variances are ideally estimated and propagated.
This limitation might also explain the superior performance of modified imputation when compared to uncertainty decoding in the context of uncertainty propagation. Both modified imputation and uncertainty decoding assume the distribution of feature uncertainty to be independent of the mixture being evaluated. However, for an uncertainty covariance tending to infinity, modified imputation converges to conventional imputation (see Equation 3.46), whereas uncertainty decoding would output a zero probability for that mixture (see Equation 3.40). This is particularly critical in the case in which a single feature or a few features of the frame exhibit a very large uncertainty. In this case uncertainty decoding would be more severely affected due to the modification of the PDFs normalizing factor.

8.1.3 Sensibility of Modified Imputation and Uncertainty Decoding to the Uncertainty Variance

When comparing ETSI AFE/RASTA ASR results in Tables 7.2 and 7.3 with their corresponding variance estimations in Figure 7.6, middle and Figure 7.7, top, it seems clear that the variance estimation accuracy has a low impact on the performance of modified imputation and uncertainty decoding.
Interestingly, the effect of employing diagonal or full covariances has opposite effects for the advanced front-end and RASTA features. Diagonal covariance uncertainty propagation led to lower error rates for non-reverberant environments in case of the RASTA features whereas the WERs for reverberant environments are slightly higher. As seen in Figure 7.6, middle and Figure 7.7, top, the use of diagonal covariance led to overestimation of the cepstra variance in the RASTA-LPCCs and to underestimation on the ETSI AFE case. That is, underestimation of covariance seems to have a positive effect in non-reverberant environments and a negative effect on reverberant environments and vice versa. As commented at the beginning of this chapter, such effects must be however circumscribed to the context of empirical measures of uncertainty combined with uncertainty decoding and modified imputation.
In general, it has been observed that the efficiency of uncertainty propagation can be sensibly increased by applying a scaling factor to the variance. However, since no theoretical justification for this optimal scaling factor was found it was not used in the final tests.

8.1.4 Improving the Advanced Front-End with Uncertainty Propagation

The front-end specified in the ETSI standard 202 050 is a complex algorithm involving multiple steps of speech enhancement and feature extraction. It is also optimized to yield very robust features. Nevertheless uncertainty propagation achieves a sensible improvement with almost no drawbacks. If we consider the diagonal propagation case combined with modified imputation (+MI+D in Section 7.2.3) the computational complexity is only of approximately twice that of the conventional features, in exchange for

a drop in the word error rates between 8% and 34%. Recognition under observation uncertainty using modified imputation also requires only a minimal modification of the conventional speech recognition process[2]. There is also no need for recursive parameter estimation like other observation uncertainty techniques [Stouten 2006].
If we take into account the transmission aspects of the standard, two additional disadvantages arise. The number of features to be transmitted through the channel are 25 instead of 13 and quantization of the variances would also be needed in a real implementation. The trade-off offered by uncertainty propagation remains nevertheless positive.

8.1.5 Posterior Propagation as MMSE-Feature Estimation

The uncertainty propagation implementation used to improve the ETSI advanced front-end simply considers the output of the speech enhancement block as 'uncertain' and described by a complex Gaussian model. The uncertainty variance is also determined in an ad hoc. fashion. For conventional implementations of MMSE estimators, however, the complex Gaussian model arises naturally. As seen in Section 5.3.2, the posterior distribution resulting from the Wiener filter for each Fourier coefficient $p(X_{kl}|Y_{kl})$ corresponds to a complex Gaussian model with a variance equal to the Bayesian error of estimation. Propagating this posterior through the feature extraction process is in fact an approximate MMSE estimation of the clean observation $\mathbf{o}$ for the Gaussian model.
Let us consider the true clean observation matrix $\mathbf{o}$ in the feature domain and the noisy Fourier spectrum matrix $\mathbf{Y}$, defined in Chapters 2 and 3, as realizations of random processes ruled by $p(\mathbf{o})$ and $p(\mathbf{Y})$. According to the definition of the MMSE estimate in Section 2.3.4 we have that

$$\hat{\mathbf{o}}^{\mathrm{MMSE}} = \arg\min_{\hat{\mathbf{o}}} \left\{ E\left\{ \|\mathbf{o} - \hat{\mathbf{o}}\|^2 \right\} \right\} = E\{\mathbf{o}|\mathbf{Y}\} = E\{\mathbf{f}(\mathbf{X})|\mathbf{Y}\}, \tag{8.2}$$

where

$$\mathbf{o} = \mathbf{f}(\mathbf{X}), \tag{8.3}$$

is a vector valued function corresponding to the feature extraction applied to $\mathbf{X}$, the original clean signal spectrum. The propagation of the Wiener filter posterior using the developed uncertainty propagation techniques (see Chapter 6) is in fact a form of approximation of the posterior

$$p(\mathbf{o}|\mathbf{Y}) = p(\mathbf{f}(\mathbf{X})|\mathbf{Y}), \tag{8.4}$$

which provides the expectation in Equation 8.2 as well as a variance of estimation. A simple example of this is the propagation of the Wiener filter posterior through the amplitude feature extraction

$$f_{kl}(\mathbf{X}) = |X_{kl}|, \tag{8.5}$$

[2]Unlike uncertainty decoding, modified imputation does not require the recomputation of the denominator in Equation 3.3 for each new frame. This denominartor is usually precomputed during training to reduce the compuational load (see HTK [Young 2006, p. 116]).

which corresponds to the posterior of the MMSE-STSA Ephraim-Malah filter.
For a more complex transformation the like the PSD based cepstra

$$o_{il} = \sum_{j=1}^{J} T_{ij} \log \left(\sum_{k=1}^{K} W_{kj} |X_{kl}|^2 \right), \tag{8.6}$$

where T_{ij} and W_{kj} correspond to the DCT and Mel-filterbank coefficients respectively (see Section 3.3.1), the transformed posterior is approximately normal distributed

$$p(\mathbf{o}_l|\mathbf{Y}_l) \approx N\left(\mathbf{o}_l; \boldsymbol{\mu}_l^{\text{CEPS}}, \boldsymbol{\Sigma}_l^{\text{CEPS}}\right), \tag{8.7}$$

where

$$\hat{\mathbf{o}}^{\text{MMSE}} \approx \boldsymbol{\mu}_l^{\text{CEPS}}, \tag{8.8}$$

and $\boldsymbol{\Sigma}_l^{\text{CEPS}}$ are obtained by applying the formulas in Sections 6.3.1, 6.2.2, 6.3.3 and 6.2.4.
The propagation of the Wiener filter posterior into the feature domain can then be seen as a method of obtaining a parametric model of speech distortion comparable with that of other observation uncertainty techniques like [Deng 2002, Benitez 2004]. This method has the additional advantage of not being limited to a single feature extraction.
More importantly, the fact that the uncertainty model is derived from a MMSE estimator in STFT domain allows this technique to make use of the existing expertise in the field of STFT speech enhancement. The extension of the additive noise framework to the suppression of reverberant noises, for example, could be carried out using [Habets 2007] with almost no modification of the presented algorithms.

8.1.6 Noise Prior conditioned by Noise Variance Estimation

The supergaussian characteristic of the developed prior was based on the distribution of the estimated noise variance. This was achieved by explicitly using the prior model of Ephraim and Cohen given in [Ephraim 2004] (see Section 4.1.2)[3].
The use of a marginalization of the variance to determine the prior can be seen as a method complementary to the one developed by Martin and used in [Martin 2002, Lotter 2005, Erkelens 2007] (see Section 4.1.1). Both methods are simplifications of a more general case, in which the full joint of the 'true' unseen Fourier coefficient given some estimated SNR parameters is computed.
Let us consider the case of a speech Fourier coefficient X estimated with a speech enhancement method dependent on ξ (i.e. Wiener filter). Let us consider also the joint distribution $p(X, \xi)$. This PDF would specify the form of the conditional $p(X|\xi)$, for all the range of values of ξ and could thus be used as a prior of X dependent on ξ.
A possible simplification of the modeling of this joint distribution is to assume that only a small range $[\xi_{\min}, \xi_{\max}]$ of values of ξ matters for the speech enhancement method and that the conditional is constant (i.e. does not change with ξ) in that range. Under such assumptions the prior distribution can be approximated as

[3]The derivation in [Erkelens 2007] uses [Ephraim 2004] implicitly as justification for the use of Gaussian mixture models although the training of the model is done as in [Martin 2002].

$$p(X) \approx p(X|\xi \in [\xi_{\min}, \xi_{\max}]), \tag{8.9}$$

which is the solution used in [Martin 2002, Lotter 2005, Erkelens 2007]. Another alternative to computing the whole joint $p(X, \xi)$ is to marginalize ξ as

$$p(X) \approx \int p(X|\xi)p(\xi)d\xi, \tag{8.10}$$

where the resulting distribution can be interpreted as the average of the conditionals for the range ξ values. If we assume $p(X|\xi)$ to be Gaussian, the model is similar to the one presented by Ephraim and Cohen.
Both models are simplifications that offer positive trade-offs for different purposes. While speech has a well known structure, 'noise' corresponds to 'everything, which is not speech'. The method presented in Section 4.4 therefore attempts to capture the supergaussian characteristics of the noise prior for a given noise estimation by averaging over various noise samples. Given that this 'noise' does not have a known structure, it does not seem feasible to find a nominal range of values for which the conditional can be easily computed. The approximation through the marginalization proposed in this work, seems therefore to offer a better trade-off for this particular purpose.

8.1.7 Limitations of Supergaussian Priors

There is a large body of evidence supporting the assumption that supergaussian priors offer a better fit to the 'true' distribution of speech as well as better properties in terms of audible quality and SNR [Martin 2002, Lotter 2005, Andrianakis 2006, Erkelens 2007, Erkelens 2007b]. Despite this fact the speech prior used in the ASR tests, computed as in [Erkelens 2007], showed a poor performance.
This is nevertheless not contradictory. The optimal trade-off between residual noise and speech attenuation is different for human listeners and ASR systems. The main objective of these tests was to show the flexibility of the purposed complex Gaussian mixture model by combining supergaussian priors for speech and noise. The use of both supergaussian models for speech and noise might require further parameter optimizations.
Traditional SNR estimation methods (see Section 2.2), are derived under the assumption of a Gaussian distributed prior. The efficiency of these methods in ASR is often improved by adjusting trade-off parameters like the parameter α controlling the recursive averaging in Equation 2.53 or Equation 2.59 or imposing non-linear upper or lower bounds to estimated parameters like $\xi_{\min}$. Most of these values are derived in an ad hoc. fashion from the experience gathered with Gaussian priors and might also already compensate the inaccuracies caused by the Gaussian assumption.

8.1.8 Supergaussian Uncertainty Propagation

The posterior distribution of each Fourier coefficient given the observable information was derived for the GMM priors in Section 5.3.4. Although not put in practice during this thesis, this posterior can thus be used for uncertainty propagation of supergaussian priors.

Since under this model, $p(X|Y)$ is a complex Gaussian mixture model and consequently $p(A|Y)$ a Rice mixture model, propagation of this supergaussian uncertainty model through STSA, PSD and filterbank transformations is similar to the Gaussian case. Propagation through other transformations is however more complicated since, the form of the distribution of uncertainty after the filterbank transformation is unknown and varies with the supergaussian characteristics of $p(X)$ and $p(D)$. During the work on this thesis various attempts were carried out to complete supergaussian propagation for the Mel-cepstral transformation through the logarithm. Unfortunately none of the attempts provided a closed form solution for this problem. The low sensibility of the algorithm to strong changes in variance estimation also discourages further investigation of supergaussian propagation until a proper use for it in the feature domain is found.

8.2 Conclusions

This thesis has analyzed the integration of statistical speech enhancement techniques in the STFT domain with an ASR system to achieve a more robust performance. This integration is achieved by replacing the conventional point estimate obtained from a speech enhancement method in the STFT domain by an 'uncertain' or probabilistic description of the clean signal which is propagated through the feature extraction into the domain of recognition. Once in this domain, observation uncertainty techniques can be employed to yield a more robust speech recognition. Consequently uncertainty propagation allows to take advantage of the expertise already existing in both STFT domain speech enhancement and observation uncertainty disciplines.

For the conventional Gaussian framework for speech enhancement, closed form solutions and approximations were developed for the obtention of an uncertain description of the clean Fourier spectrum estimated with the Wiener, MMSE-STSA and MMSE-LSA estimators based on complex Gaussian models. An empirical method for the computation of uncertainty for speech enhancement methods, which combines multiple techniques, was also proposed. This was illustrated using the ETSI AFE speech enhancement system.

The propagation of both MMSE estimation error and empirical measures of uncertainty for the ETSI AFE proved to be an effective method of improving recognition rates with very few drawbacks. Tests results included a notable improvement of the ETSI AFE baseline on the AURORA5 database, which is a highly demanding benchmark. The conventional Gaussian framework for speech enhancement was also extended to a complex Gaussian mixture model. A method for the estimation of the noise prior parameters given a noise variance estimation was also presented. MMSE estimators of the short-time spectral amplitude and log-spectral amplitude derived under this model improved their Gaussian counterparts, although only the MMSE-LSA estimator was better in all scenarios.

The propagation of the complex Gaussian model through the Mel-cepstral, RASTA-LPCC and ETSI AFE transformations has been studied. The piecewise uncertainty propagation has been proposed as a solution for the problem of uncertainty propagation. A set of exact closed form solutions, analytical approximations and pseudo-Montecarlo

implementations have been used to solve different propagation problems through non-linear transformations, attaining a high accuracy of propagation under multiple conditions. Such solutions can also be combined to yield uncertainty propagation algorithms for other feature extractions.

8.3 Future Work

Future work will explore some of the research directions pointed out in the discussion. Improving the accuracy of uncertainty propagation furthermore or achieving efficient supergaussian propagation are possible objectives. From a theoretical point of view, the perspective of fully capturing the uncertainty derived from the Gaussian model for speech enhancement seems particularly interesting. Incorporating issues like the incorrectness of the assumption of ergodicity or, in general, errors in the estimation of SNR, would for sure improve the posterior propagation technique presented. One particular aspect to be improved is the way uncertainty is integrated with the ASR system. There is currently plenty of research on this field (see [Liao 2008] for a review of the techniques). The combination of uncertainty propagation with improved versions of uncertainty decoding, like for example [Umbach 2006], might yield substantial improvements.

List of Abbreviations

AFE	Advanced Front-End.
ASR	Automatic Speech Recognition.
Bark	Bark frequency.
BE	Blind Equalization.
CEPS	CEPStra.
CMS	Cepstral Mean Subtraction.
D	Diagonal covariance (by uncertainty propagation).
DCT	Discrete Cosine Transform.
DFT	Discrete Fourier Transform.
DSR	Distributed Speech Recognition.
EM	Expectation-Maximization.
ETSI	European Telecommunications Standards Institute.
EXP	EXPonential.
F	Full covariance (by uncertainty propagation).
FIR	Finite Impulse Response.
GMM	Gaussian Mixture Model.
GN	Gaussian Noise.
GS	Gaussian Speech.
HMM	Hidden Markov Model.
HTK	Hidden Markov model ToolKit.
ICA	Independent Component Analysis.
IDFT	Inverse Discrete Fourier Transform.
IIR	Infinite Impulse Response.
IMCRA	Improved Minima Controled Recursive Averaging.
ISTFT	Inverse Short-Time Fourier Transform.
LMMSE	Linear Minimum Mean Square Error.
LogE	Log-Energy.
LOG	LOGarithm.
LPC	Linear Prediction Coefficient.
LPCC	Linear Prediction Coefficient Cepstra.
LSA	Log-Spectral Amplitude.
MAP	Maximum A Posteriori.
Mel	Mel frequency.
MFCC	Mel-Frequency Cesptral Coefficient.
MI	Modified Imputation.
ML	Maximum Likelihood.
MMSE	Minimum Mean Square Error.
MSE	Mean Square Error.
OM	Optimally Modified.
PDF	Probability Density Function.
PLP	Perceptual Linear Prediction.

PMC	Parallel Model Combination.
PSD	Power Spectral Density.
RASTA	RelAtive SpecTrAl.
RiMM	Rice Mixture Model.
RMM	Rayleigh Mixture Model.
SN	Supergaussian Noise.
SNR	Signal to Noise Ratio.
SPLICE	Stereo based Piecewise Linear Compensation for Environments.
SS	Supergaussian Speech.
STFT	Short-Time Fourier Transform.
STSA	Short-Time Spectral Amplitude.
UD	Uncertainty Decoding.
VAD	Voice Activity Detection.
VTS	Vector Taylor Series.
W	Wiener filter.
WA	Word Accuracy.
WC	Word Correctness.
WER	Word Error Rate.

List of Symbols

$*^T$	Transpose operator.
$\mathbf{A}_l$	Frame of the STSA of $x(t)$ (K elements).
A_{il}	Acceleration feature.
A_{kl}, A	Amplitude of X_{kl}, X (STSA).
$\mathbf{B}_l$	Frame of Bark-filterbank features (J elements).
B_{jl}	Bark-filterbank feature.
C_{il}	Cepstral feature.
C_i^{ref}	Reference of the flat spectrum cepstra.
$\text{Cov}\{\}$	Covariance of a random variable.
$\text{Cum}_n\{\}$	Cumulant of order n of a random variable.
$\tilde{C}_{il}$	Cepstral feature after blind equalization.
D	Number of deletions found in a transcription.
D_I	Imaginary component of the Fourier coefficient D.
D_R	Real component of the Fourier coefficient D.
D_{il}	Delta feature.
D_{kl}, D	Fourier coefficient of $d(t)$.
$E\{\}$	Expectation of a random variable.
E_l	Log-energy feature.
$\mathbf{F}^{\mathbf{I}}$	Matrix of imaginary components of each element of $\mathbf{F}$.
$\mathbf{F}^{\mathbf{R}}$	Matrix of real components of each element of $\mathbf{F}$.
$\mathbf{F}$	IDFT matrix.
$F_{kt'}$	Element of the IDFT matrix.

G_{kl}	Gain function in STFT domain.
$\mathbf{H}_l$	Frame of perceptually modified features (J elements).
$H(z)$	RASTA filter gain function in z domain.
H_0	Hypothesis of speech absence.
H_1	Hypothesis of speech presence.
H_{jl}	Perceptually modified power spectrum feature.
I	Number of cepstral coefficients.
I	Number mixtures in the GMM prior of X.
I	Number of insertions found in a transcription.
I_0, I_1	Modified Bessel functions of the first and second kind.
Im{}	Imaginary component of a complex number.
J	Number of Mel/Bark filters in filterbank .
J	Number of mixtures in the GMM prior of D.
J	Number of states in a HMM.
K	Number of frequency bins up to half the sampling frequency.
$K3^*_{kl}$	Cumulant of order three of $*$ (i.e. PSD, Mel, CEPS).
$K4^*_{kl}$	Cumulant of order four of $*$ (i.e. PSD, Mel, CEPS).
Kurt^*	Kurtosis of $*$ (i.e. PSD, Mel, CEPS).
L	Total number of frames of the STFT.
$L_n^\alpha()$	Laguerre polynomial ($\alpha = 0$ and is omitted).
$\mathbf{L}_l$	Frame of uncertain features in logarithm domain (J elements).
$\mathbf{M}_l$	Frame of Mel-filterbank features (J elements).
M	Number of samples shifted between STFT frames.
M	Total number of mixtures in the GMM emitting probability of a HMM.
$M_X(\mu)$	Moment generating function of the random variable X.
M_{jl}	Mel-filterbank feature.
N	Number of samples per STFT frames.
N	Number of samples in a Montecarlo simulation.
N	Total number of words in a transcription.
N	Number of Sigma points in the unscented transform.
$N()$	Normal or Gaussian distribution.
N_W	Total number of words in a HMM.
$\mathbf{P}_l$	Frame of PSD features (K elements).
P_{kl}	PSD of X_{kl}.
R_{kl}, R	Amplitude of Y_{kl}, Y.
Re{}	Real component of a complex number.
S	Number of substitutions found in a transcription.
S_i, S_j	Two different states of a HMM.
S_i	Sigma point of the unscented transform.
$\mathbf{T}$	Linear transformation matrix.
T_{ij}	Element of $\mathbf{T}$.
V_{jk}	Weight of the Bark-filterbank.
Var{}	Variance of a random variable.

W_i	Unscented transform weights.
W_{jk}	Weight of the Mel-filterbank.
W_{jm}	Weight of the GMM emitting probability of a HMM.
$\mathbf{X}$	STFT Matrix of $x(t)$ ($K \cdot L$ Fourier coefficients).
$\mathbf{X}_l$	Frame of the STFT of $x(t)$ (K Fourier coefficients).
$\mathbf{X}'_l$	Frame of the STFT of $x(t)$ (N Fourier coefficients, including those above half the sampling frequency).
$\mathbf{X}'^R_l$	Frame of real components of $\mathbf{X}'_l$.
$\mathbf{X}'^I_l$	Frame of imaginary components of $\mathbf{X}'_l$.
X^*	Complex conjugate of the Fourier coefficient X.
X_I	Imaginary component of the Fourier coefficient X.
X_R	Real component of the Fourier coefficient X.
X_{kl}, X	Fourier coefficient of $x(t)$.
$\hat{X}^N_{kl}$	Fourier coefficient of $\hat{x}_N(t)$.
$\hat{X}^W_{kl}$	Fourier coefficient of $\hat{x}_W(t)$.
$\mathbf{Y}$	STFT Matrix of $y(t)$ ($K \cdot L$ Fourier coefficients).
$\mathbf{Y}_l$	Frame of the STFT of $y(t)$ (K Fourier coefficients).
Y_I	Imaginary component of the Fourier coefficient Y.
Y_R	Real component of the Fourier coefficient Y.
Y_{kl}, Y	Fourier coefficient of $y(t)$.
$\Gamma()$	Gamma function.
Ω_{ij}	Mixture weight of the RiMM clean speech posterior.
$\Phi()$	Confluent hypergeometric function.
$\Sigma^*_{kk'l}$, $\Sigma^*_{jj'l}$, $\Sigma^*_{ii'l}$	Covariance between features of $*$ (i.e. PSD, Mel, CEPS).
$\mathbf{\Sigma}^o_l$	Covariance matrix of $p(\mathbf{o}_l)$.
$\mathbf{\Sigma}_{jm}$	Covariance matrix of mixture m belonging to state S_j.
$\mathbf{\Sigma}^y_l$	Covariance matrix of the input PDF for the unscented transform.
Θ	Weight of each frame in the delta features.
α	Phase of Fourier coefficient X.
α	Forgetting factor for the noise recursive averaging.
α	Forgetting factor for the decision directed ξ estimation.
α	Pre-emphasis coefficient.
α_{jmi}	Weighting factor in modified imputation.
$\tilde{\alpha}$	Forgetting factor computed from speech probability p.
ϑ_{kl}, ϑ	Phase of Fourier coefficient Y_{kl}, Y.
$\delta_l(j)$	Maximum joint probability at time step l and state S_j.
δ_{kl}	Change between AFE steps $x_N(t)$ and $x_W(t)$.
γ_{kl}, γ	A posteriori SNR of noisy Fourier coefficient Y_{kl}, Y.
γ_j	Mixture a posteriori SNR in the GMM speech enhancement model.
κ	Parameter of the unscented transform controlling higher order information.
λ	Variance of the uncertain Fourier coefficient X.

λ	Variance of the Wiener filter posterior.
λ_{D_j}	Mixture variance in the GMM prior of D.
$\tilde{\boldsymbol{\lambda}}$	Matrix of variance estimations $\tilde{\lambda}_{kl}$ ($K \cdot L$ elements).
$\lambda_{D_{kl}}$, λ_D	Variance of D_{kl}, D.
λ_{X_i}	Mixture variance in the GMM prior of X.
$\lambda_{X_{kl}^N}$	Variance of X_{kl}^N.
$\lambda_{X_{kl}}$, λ_X	Variance of X_{kl}, X.
$\lambda_{\delta_{kl}}$	Variance of δ_{kl}.
λ_{ij}	Mixture variance of the complex GMM posterior.
$\tilde{\lambda}$, $\tilde{\lambda}_{kl}$	Estimation of λ, λ_{kl}.
Ei()	Exponential integral function.
μ_{kl}^*, μ_{jl}^*, μ_{il}^*	Mean of feature $*$ of frame l (i.e. PSD, Mel, CEPS).
μ_I	Imaginary component of μ_X.
μ_R	Real component of μ_X.
μ_l	Stepsize of the blind equalization.
$\boldsymbol{\mu}_l^o$	Mean vector of $p(\mathbf{o}_l)$.
$\boldsymbol{\mu}_{jm}$	Mean matrix of mixture m belonging to state S_j .
$\boldsymbol{\mu}_l^y$	Mean vector of the input PDF for the unscented transform.
μ_X	Mean of the complex Gaussian random variable X.
ν	Fixed parameter of the blind equalization.
ν, ν_{kl}	Rice ratio of the posterior $p(A_{kl}\|Y_{kl})$.
ψ_j	Pre-emphasis multiplicative gain.
$\psi_l(j)$	Variable tracking the occupied state sequence in Viterbi decoding.
$\sigma_{D_j}^2$	Mixture variance in the GMM prior of D_R, D_I.
σ_D^2	Variance of D_R, D_I.
$\sigma_{X_i}^2$	Mixture variance in the GMM prior of X_R, X_I.
σ_X^2	Variance of X_R, X_I.
θ	Phase of μ_X.
θ	Length of the window in the computation of Delta features.
$\xi_{\max}$	Maximum value of ξ considered to determine the speech prior.
$\xi_{\min}$	Floor value for ξ estimation.
$\xi_{\min}$	Minimum value of ξ considered to determine the speech prior.
ξ_{ij}	Mixture a priori SNR in the GMM speech enhancement model.
ξ_{kl}, ξ	A priori SNR of noisy Fourier coefficient Y_{kl}, Y.
a_{ij}	State transition probability.
a_1	Normalized feedback coefficient of RASTA filter.
b_j	State emission probability.
b_{il}	Bias of the blind equalization.
b_d	Normalized feedforward coefficients of RASTA filter.
c	Euler-Mascheroni constant.
$d(t)$	Noise signal in time domain.
d	Frame shift in FIR and IIR filters.

f^{Mel}	Mel frequency.
f^{Bark}	Bark frequency.
$\mathbf{g}()$	Non-linear vector valued feature extraction function.
$\mathbf{f}()$	Non-linear matrix valued feature extraction function.
$f_{kl}()$	Element of the matrix output provided by $\mathbf{f}()$.
$h(t')$	Window function of the STFT.
i	Imaginary unit.
i	Index of HMM state.
i	Index of cepstral coefficients.
i	Index of mixture in the GMM prior of X.
j	Index of HMM state.
j	Index of Mel/Bark filterbank.
j	Index of mixture in the GMM prior of D.
k	Index of frequency bins within a STFT frame.
l	Index of STFT frame.
m	Index of mixture in the GMM of a HMM.
$\mathbf{o}$	Matrix of features (observation).
$\mathbf{o}_l$	Frame of features (observation).
$\mathbf{o}_l^{jm}$	Estimated clean frame of features for mixture m of state S_j.
p	Probability of speech presence $p(H_1\|Y)$.
$p()$	PDF of a random variable.
$\hat{p}()$	Probability estimated using uncertainty decoding.
q_l	Occupied state at time step l.
$\mathbf{q}$	Sequence of occupied states at time steps $l = 1 \cdots L$.
r	Ratio of the Rice distribution.
r_{D_j}	Variance ratio in the GMM prior of D.
r_{X_i}	Variance ratio in the GMM prior of X.
t	Index of time domain samples.
t'	Index of time domain samples within one STFT frame.
w_{D_j}	Mixture weight in the GMM prior of D.
w_{X_i}	Mixture weight in the GMM prior of X.
$\mathbf{x}$	Random vector variable.
$x(t)$	Clean signal in time domain.
$\tilde{x}(t)$	Clean signal after pre-emphasis.
x_{jl-d}	Logarithm domain inputs to the RASTA filter.
$x_{t'l}$	Time domain sample of $x(t)$ within one STFT frame.
$\mathbf{x}_l$	Frame of N time domain samples.
$\hat{x}_N(t)$	Estimated clean signal after the noise reduction block of AFE.
$\hat{x}_W(t)$	Estimated clean signal after the waveform processing block of AFE.
$y(t)$	Noisy signal in time domain.
y_{jl}	Output of RASTA filter.
$y_{t'l}$	Time domain sample of $y(t)$ within one STFT frame.
$\mathbf{y}_l$	Frame of N time domain samples.

Bibliography

[Acero 2000] Alex Acero, Li Deng, Trausti Kristjansson, and Jerry Zhang. HMM adaptation using vector Taylor series for noisy speech recognition. Technical report, Microsoft Research, 2000.

[Agarwal 1999] Anshu Agarwal and Yan Ming Cheng. Two-stage mel-warped wiener filter for robust speech recognition. In *Proc. IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 12–15, 1999.

[Andrianakis 2006] I. Andrianakis and P.R. White. MMSE speech spectral amplitude estimators with chi and gamma speech priors. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 3, pages III–III, May 2006.

[Benitez 2004] M. C. Benítez, J. C. Segura, A. Torre, J. Ramírez, and A. Rubio. Including uncertainty of speech observations in robust speech recognition. In *Proc. International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 137–140, 2004.

[Boll 1979] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 27(2):113–120, Apr 1979.

[Cohen 2003] I. Cohen. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Trans. on Speech and Audio Processing*, 11(5):466–475, Sept. 2003.

[Cohen 2001] Israel Cohen and Baruch Berdugo. Speech enhancement for non-stationary noise environments. *Signal Processing*, 81(11):2403 – 2418, 2001.

[Cooke 2001] Martin Cooke, Phil Green, Ljubomir Josifovski, and Ascension Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34(3):267 – 285, 2001.

[Davis 1980] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech, Signal Processing*, 28 (4)(2):357– 366, Apr 1980.

[Deller 1987] J. R. Deller, J. H. L. Hansen, and J. G. Proakis. *Discrete-Time Processing of Speech Signals*. Prentice-Hall, Inc., 1987.

[Dempster 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Royal Statistical Society, Series B (Methodological)*, 39 (1):1–38, 1977.

[Deng 2002] Li Deng, Jasha Droppo, and Alex Acero. Exploiting variances in robust feature extraction based on a parametric model of speech distortion. In *Proc. International Conference on Spoken Language Processing (ICSLP)*, 2002.

[Droppo 2002] J. Droppo, A. Acero, and Li Deng. Uncertainty decoding with splice for noise robust speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, volume 1, pages I–57–I–60 vol.1, 2002.

[Ellis 2005] Daniel P. W. Ellis. PLP and RASTA (and MFCC, and inversion) in Matlab, 2005. online web resource.

[Ephraim 2004] Y. Ephraim and I. Cohen. *Recent Advancements in Speech Enhancement*, pages 1–22. CRC Press, May 17, 2004.

[Ephraim 1984] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, 32(6):1109–1121, Dec 1984.

[Ephraim 1985] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean square error log-spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-33:443–445, 1985.

[Erkelens 2007] J. S. Erkelens, J. Jensen, and R. Heusdens. Speech enhancement based on rayleigh mixture modeling of speech spectral amplitude distributions. In *Proc. European Signal Processing Conference (EUSIPCO)*, pages 65–69, September 2007.

[Erkelens 2007b] J.S. Erkelens, R.C. Hendriks, R. Heusdens, and J. Jensen. Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors. *IEEE Trans. on Audio, Speech, and Language Processing*, 15(6):1741–1752, Aug. 2007.

[ETSI 2007] ETSI. *ETSI Standard document, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech recognition; Front-end feature extraction algorithm; Compression algorithms, ETSI ES 202 050 v1.1.5 (2007-01)*, January 2007.

[Gales 1995] M. J. F. Gales. *Model-Based technique for noise robust speech recognition.* PhD thesis, Gonville and Caius College, 1995.

[Gradshteyn 2007] I. S. Gradshteyn and I.M Ryzhik. *Table of Integrals, Series and Products.* Elsevier, 2007.

[Gudbjartsson 1995] H. Gudbjartsson and S. Patz. The rician distribution of noisy MRI data. *Magnetic resonance in medicine*, 34(6):910–914, 1995.

[Habets 2007] Emanuël Anco Peter Habets. *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement.* PhD thesis, Technische Universiteit Eindhoven, 2007.

[Hermansky 1985] H. Hermansky, B. Hanson, and H. Wakita. Perceptually based linear predictive analysis of speech. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 1985.

[Hermansky 1994] H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Trans. on Speech and Audio Processing*, 2(4):578–589, Oct 1994.

[Hirsch 2002] Guenter Hirsch. *Experimental Framework for the Performance Evaluation of Speech Recognition Front-ends on a Large Vocabulary Task*. Niederrhein University of Applied Sciences, November 2002.

[Hirsch 2007] Guenter Hirsch. *Aurora-5 Experimental Framework for the Performance Evaluation of Speech Recognition in Case of a Hands-free Speech Input in Noisy Environments*. Niederrhein University of Applied Sciences, August 2007.

[Hirsch 2000] H. G. Hirsch and D. Pearce. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proc. Automatic Speech Recognition: Challenges for the new Millenium (ASR)*, 2000.

[Hirsch 2005] H.G. Hirsch and H. Finster. The simulation of realistic acoustic input scenarios for speech recognition systems. In *Proc. Interspeech*, 2005.

[Umbach 2006] Valentin Ion and Reinhold Haeb-Umbach. Improved source modeling and predictive classification for channel robust speech recognition. In *Proc. Interspeech*, 2006.

[Itahashi 1990] Shuichi Itahashi. Recent speech database projects in japan. In *Proc. International Conference on Spoken Language Processing (ICSLP)*, pages 1081–1084, 1990.

[Jaynes 1970] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 1970.

[Johnson 1970] Norman L. Johnson. *Continous Univariate Distributions -1*. Wiley Interscience, 1970.

[Johnson 1970b] Norman L. Johnson. *Continous Univariate Distributions -2*. Wiley Interscience, 1970.

[Kay 1993] Steven M. Kay. *Fundamentals of Statistical Signal Processing*. Prentice Hall signal processing series, 1993.

[Kolossa 2005] D. Kolossa, A. Klimas, and R. Orglmeister. Separation and robust recognition of noisy, convolutive speech mixtures using time-frequency masking and missing data techniques. In *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 82–85, Oct. 2005.

[Kolossa 2006] D. Kolossa, H. Sawada, R.F. Astudillo, R. Orglmeister, and S. Makino. Recognition of convolutive speech mixtures by missing feature techniques for ICA. In *Proc. Asilomar Conference on Signals, Systems, and Computers*, pages 1397–1401, 2006.

[Kolossa 2008] Dorothea Kolossa. *Independent Component Analysis for Environmentally Robust Speech Recognition*. PhD thesis, Berlin Institute of Technology, 2008.

[Kristjansson 2002] Trausti T. Kristjansson and Brendan J. Frey. Accounting for uncertainity in observations: a new paradigm for robust automatic speech recognition. Technical report, Probabilistic and Statistical Inference Group, University of Toronto, 2002.

[Kuroiwa 2003] S. Kuroiwa, S. Tsuge, and F. Ren. Blind equalization via minimization of VQ distortion for ETSI standard DSR front-end. In *Proc. International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*, pages 585–590, Oct. 2003.

[TIDigits 1993] LDC. Tidigits speech database: Studio quality speaker-independent connected-digit corpus, 1993.

[Liao 2008] H. Liao and M.J.F. Gales. Issues with uncertainty decoding for noise robust automatic speech recognition. *Speech Communication*, 50(4):265 – 277, 2008.

[Lotter 2005] Thomas Lotter and Peter Vary. Speech enhancement by map spectral amplitude estimation using a super-gaussian speechmodel. *EURASIP Journal on Applied Signal Processing*, 7:1110–1126, 2005.

[Macho 2001] D. Macho and Yan Ming Cheng. SNR-dependent waveform processing for improving the robustness of asr front-end. In *Proc. International Conference on Acoustics Speech and Signal Processing (ICASSP)*, volume 1, pages 305–308 vol.1, 2001.

[Macho 2002] D. Macho, L. Mauuary, B. Noé, Y. M. Cheng, D. Ealey, D. Jouvet, H. Kelleher, D. Pearce, and F. Saadoun. Evaluation of a noise-robust DSR front-end on aurora databases. In *Proc. International Conference on Spoken Language Processing (ICSLP)*, September 2002.

[Mandelartz 2009] Philipp Mandelartz. Robust feature extraction based on the etsi distributed speech recognition standard. Master's thesis, Berlin Institute of Technology, 2009.

[Markel 1967] J. D. Markel and A. H. Gray Jr. *Linear Prediction of Speech*. Springer-Verlag, 1967.

[Martin 2001] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. on Speech and Audio Processing*, 9(5):504–512, Jul 2001.

[Martin 2002] R. Martin. Speech enhancement using mmse short time spectral estimation with gamma distributed speech priors. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 1, pages I–253–I–256 vol.1, 2002.

[Martin 2005] R. Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Trans. on Speech and Audio Processing*, 13(5):845–856, Sept. 2005.

[Martin 2004] R. Martin, D. Malah, V. R. Cox, and A. J. Accardi. A noise reduction preprocessor for mobile voice communication. *EURASIP Journal on Applied Signal Processing*, 8:1046–1058, 2004.

[McAulay 1980] R. McAulay and M. Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. Acoust., Speech, Signal Processing*, 28(2):137–145, Apr. 1980.

[Middelton 1960] D. Middelton. *Introduction to Statistical Communication Theory.* McGraw-Hill, 1960.

[Chrysostomos 1993] Chrysostomos L. Nikias and Athina P. Petropulu. *Higher-order spectra analysis, a nonlinear signal processing framework.* Prentice Hall signal processing series, 1993.

[Pearlman 1978] William A. Pearlman and Robert M. Gray. Source coding of the discrete fourier transform. *IEEE Trans. in Information Theory*, IT-24 (6):683–692, 1978.

[Porter 1984] J. Porter and S. Boll. Optimal estimators for spectral restoration of noisy speech. *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 9:53–56, Mar 1984.

[Press 2007] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes: The Art of Scientific Computing, Third Edition.* Cambridge University Press, 2007.

[Proakis 2007] John G. Proakis and Dimitris G. Manolakis. *Digital Signal Processing.* Prentice Hall, 2007.

[Quackenbush 1988] S. Quackenbush, T. Barnwell, and M. Clements. *Objective Measures of Speech Quality.* Englewood Cliffs, NJ: Prentice-Hall, 1988.

[Rabiner 1989] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *IEEE Proceedings, Vol.77 No.2*, February 1989.

[Rice 1944] S. O. Rice. *Mathematical Analysis of Random Noise*, volume 23. Bell Telephone Labs Inc., 1944.

[Shannon 2003] Ben J. Shannon and Kuldip K. Paliwal. A comparative study of filter bank spacing for speech recognition. In *Proc. Microelectronic Engineering Research Conference*, 2003.

[Srinivasan 2007] S. Srinivasan and D. Wang. Transforming binary uncertainties for robust speech recognition. *IEEE Trans. Audio, Speech and Language Processing*, 15(7):2130–2140, September 2007.

[Srinivasan 2006] S. Srinivasan and DeLiang Wang. A supervised learning approach to uncertainty decoding for robust speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages I –I, 14-19 2006.

[Stouten 2006b] V. Stouten, H. Van Hamme, and P. Wambacq. Application of minimum statistics and minima controlled recursive averaging methods to estimate a cepstral noise model for robust ASR. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 1, pages I–I, May 2006.

[Stouten 2006] V. Stouten, H. Van hamme, and W. Wambacq. Model based feature enhancement with uncertainty decoding for noise robust ASR. *Speech Communication.*, 48(11):1502–1514, 2006.

[Tchorz 2000] J. Tchorz. *Auditory-based signal processing for noise suppression and robust speech recognition.* PhD thesis, Oldenburg University, 2000.

[Julier 1996] Julier S.J. und Uhlmann J.K. A general method for approximating nonlinear transformations of probability distributions. Technical report, Dept. of Engineering Science, University of Oxford, Oxford, UK, 1996.

[Vergin 1995] R. Vergin and D. O'Shaughnessy. Pre-emphasis and speech recognition. In *Canadian Conference on Electrical and Computer Engineering*, volume 2, pages 1062–1065 vol.2, Sep 1995.

[Wang 1982] D. Jae Lim Wang. The unimportance of phase in speech enhancement. *IEEE Trans. Acoust., Speech, Signal Processing*, 30 (4)(4):679– 681, Aug 1982.

[Yoma 1998] N.B. Yoma, F. McInnes, and M. Jack. Improving performance of spectral subtraction in speech recognition using a model for additive noise. *IEEE Trans. Speech, Audio Processing*, 6 (6):579–582, 1998.

[Yoma 2002] N.B. Yoma and M. Villar. Speaker verification in noise using a stochastic version of the weighted viterbi algorithm. *IEEE Trans. Speech, Audio Processing*, 10 (3):158–166, 2002.

[Young 2006] S. Young. *The HTK Book (for HTK Version 3.4).* Cambridge University Engineering Department., 2006.

[Young 1989] S.J. Young, N.H. Russell, and J.H.S. Thornton. Token passing: a simple conceptual model for connected speech recognition systems. Technical Report CUED/FINFENG /TR.38, Cambridge University Engineering Department, 1989.