Chapter 3
Uncertainty Propagation

Ramón Fernandez Astudillo, Dorothea Kolossa

Abstract While it is often fairly straightforward to estimate the reliability of speech features in the time-frequency domain, this may not be true in other domains more amenable to speech recognition, such as e.g. for RASTA-PLP features or those obtained with the ETSI advanced front end. In such cases, one useful approach is to estimate the uncertainties in the domain where noise-reduction preprocessing is carried out, and to subsequently transform the uncertainties, along with the actual features, to the recognition domain. In order to develop suitable approaches, we will first give a short overview of relevant strategies for propagating probability distributions through nonlinearities. Secondly, for some feature domains suitable for robust recognition, we will show possible implementations and sensible approximations of uncertainty propagation and discuss the associated error margins and trade-offs.

3.1 Uncertainty Propagation

In automatic speech recognition (ASR) an incoming speech signal is transformed into a set of observation vectors $\mathbf{x} = \mathbf{x}_1 \cdots \mathbf{x}_L$ which are input to the recognition model. ASR systems which work with uncertain input data replace this observation set by a distribution of possible observed values conditioned on the available information $\mathbf{I}$, $p(\mathbf{x}_1 \cdots \mathbf{x}_L|\mathbf{I})$. This uncertainty represents the lost information that causes the mismatch between the observed speech and the trained ASR model. By combining this distribution with observation uncertainty techniques [23], superior recognition robustness can be attained. There are multiple sources from which this

Ramón Fernandez Astudillo
Electronics and Medical Signal Processing Group, TU Berlin, 10587 Berlin, e-mail: ramon@astudillo.com

Dorothea Kolossa
Electronics and Medical Signal Processing Group, TU Berlin, 10587 Berlin, e-mail: dorothea.kolossa@gmx.de

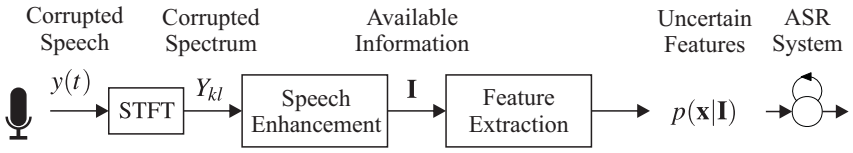

Fig. 3.1: Speech recognition using speech enhancement in short-time Fourier domain and uncertainty propagation. The available information obtained after speech enhancement in the STFT domain $\mathbf{I}$ is transformed into the feature domain to yield a probabilistic or uncertain description of the features $p(\mathbf{x}|\mathbf{I})$.

distribution over the features can be computed. Uncertainty of observation has been determined directly from speech enhancement techniques in the feature domain, like SPLICE [10], model based feature enhancement (MBFE) [31] or by the use of a state-space noise model [33]. Other models of uncertainty have been obtained by considering the relation between speech and additive noise in the logarithm domain, a step common to many feature extractions. Using this relation, uncertainty has been computed based on parametric models of speech distortion [9, 34], or the residual uncertainty after feature enhancement [5]. Other approaches have exploited different sources of uncertainty like channel transmission errors [17] or reverberation, as described in detail in Chapter 9.

This chapter discusses a particular set of techniques, which aim to determine $p(\mathbf{x}_1 \cdots \mathbf{x}_L|\mathbf{I})$ by propagating the information attained from pre-processing in the short-time Fourier transform (STFT) domain into the feature domain used for recognition (see Fig. 3.1). The STFT domain is the domain most often used for single-channel and multi-channel speech enhancement. Speech distortion phenomena like additive noise, reverberation, channel distortions and cocktail party effect can be modeled more easily in this domain than in the domains where speech recognition takes place. Some speech characteristics like vocal tract transfer function or fundamental frequency are also easier to model in this domain. Furthermore, the time domain signal can be completely recovered from the STFT of the signal, which makes this domain particularly suitable for speech enhancement targeted for human listeners.

Despite the positive properties of the STFT domain, it is rarely used for automatic speech recognition. Non-linear transformations of the STFT like the Mel frequency cepstral coefficients (MFCCs) or the relative spectral filtered perceptual linear prediction features (RASTA-PLP) are used instead. Such features have a smaller intra and inter speaker variability and achieve a more compact and classifiable representation of the acoustic space. The use of STFT domain speech enhancement is therefore usually limited to obtaining a point-estimate of the clean speech which is transformed into the domain of recognition features.

Uncertainty propagation techniques attempt to increase the amount of information which is passed from the speech enhancement step, usually in the STFT domain, to the domain of recognition features by adding a measure of reliability. This idea was first used in the context of source separation by means of independent com-

ponent analysis (ICA) [20]. For this purpose, the estimated STFT of each speaker was treated as a Gaussian variable. The uncertainty was then used to compensate for the loss of information incurred by time-frequency post-masking of ICA results. This same model was later employed to propagate the uncertainty generated after single channel speech enhancement, using the estimated noise variance [21]. These models were later extended to the complex Gaussian uncertainty model in [3], which arises naturally from the use of conventional minimum mean square error (MMSE) speech enhancement techniques as a measure of uncertainty [4]. In both cases, the transformation of the uncertain described features into the domain of recognition features was attained by using a piecewise approach combining closed form solutions with pseudo-Monte Carlo approximations.

Similarities can be drawn between these two techniques and the works in [28,29]. Here, the STFT domain was reconstructed using missing feature techniques [26], which allows the computation of an MMSE estimate of the unreliable components from the reliable components as well as a variance or measure of uncertainty. This variance was then propagated into the domain of recognition features by using general non-linear approximators like multilayer perceptrons [28] and regression trees [29], which had to be previously trained using noisy data samples. Other work which can also be related to uncertainty propagation techniques is that in [30]. Here, a noise variance estimate in the STFT domain was obtained with a minimum statistics based noise variance estimator, typically used for speech enhancement. This estimate along with a measure of reliability of its estimation was propagated into the MFCC domain using the log-normal assumption [14] and combined with MBFE techniques [30].

This chapter will be centered on the piecewise propagation approach, and it is organized as follows: Section two deals with the problem of modeling uncertainty in the spectral domain, Section three with the propagation, Section four introduces some results and draws conclusions.

3.2 Modeling Uncertainty in the STFT Domain

Let $y(t)$ be a corrupted version of a clean speech signal $x(t)$. A time frequency representation of $y(t)$ can be obtained using an N point STFT as

$$Y_{kl} = \sum_{t'=1}^{N} y\left(t' + (l-1)M\right) h(t') \exp\left(-2\pi j \frac{(t'-1)(k-1)}{N}\right), \tag{3.1}$$

where $t' \in [1 \cdots N]$ is the sample index within the frame, $l \in [1 \cdots L]$ is the frame index, $k \in [1 \cdots K]$ corresponds to the index of each Fourier bin up to half the sampling frequency and j is the imaginary unit. Each frame is shifted by M samples with respect to the previous and weighted by a window function $h(t')$.

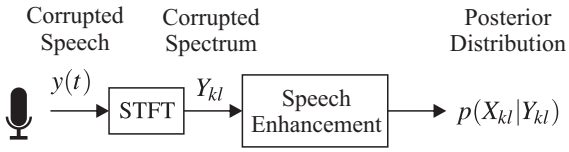

Fig. 3.2: The available information after speech enhancement in the STFT domain is summarized in form of a posterior distribution.

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

$$\hat{X}_{kl} = G_{kl} \cdot Y_{kl}. \tag{3.2}$$

This gain can be obtained, for example, from multi-channel speech enhancement, like ICA post-masking [20], missing feature techniques [29] or conventional MMSE speech enhancement [11]. Apart from this point-estimate, uncertainty propagation algorithms need of a reliability measure in form of a variance defined as

$$\lambda_{kl} = E\left\{\left|X_{kl} - \hat{X}_{kl}\right|^2\right\}. \tag{3.3}$$

where X_{kl} is the unknown clean Fourier coefficient. When MMSE estimations are used, a measure of reliability can be easily obtained from the variance of the resulting posterior distribution [4, 29]. In cases in which more complex algorithms are used, like in the case of ICA with post-masking [20] or the ETSI advanced front-end [13], approximate measures of reliability can be obtained based on the amount of change inflicted by the speech enhancement algorithm [2, 20].
Once an estimation of the clean Fourier coefficient $\hat{X}_{kl}$ and a corresponding measure of reliability λ_{kl} have been obtained, this information can be used to yield a probabilistic description of each clean Fourier coefficient X prior to propagation (see Fig. 3.2). Although it would also be possible to include the cross-covariance between different Fourier coefficients, this is usually ignored for simplicity. This is a usual simplification also used when employing similar models for speech enhancement [12, 24].
With regard to the type of distributions used to model each uncertain Fourier coefficient, in [20], a Gaussian distribution with variance equal to the uncertainty variance was used as

$$|X_{kl}| = \left|\left|\hat{X}_{kl}\right| + \delta_{kl}\right| \tag{3.4}$$

where δ_{kl} followed the zero mean Gaussian distribution

$$p(\delta_{kl}) = \frac{1}{\sqrt{2\pi\lambda_{kl}}} \exp\left(-\frac{\delta_{kl}^2}{2\lambda_{kl}}\right). \tag{3.5}$$

In [28, 29] the uncertainty was determined using a MMSE estimation based on missing feature techniques [26]. The posterior of this MMSE estimator resulted in a Gaussian mixture model for each Fourier coefficient amplitude[1].

The works in [2, 3] present an extension of the Gaussian model in [20], in which the distribution used is circularly symmetric complex Gaussian

$$p(X_{kl}|Y_{kl}) = \frac{1}{\pi\lambda_{kl}} \exp\left(-\frac{|X_{kl} - \hat{X}_{kl}|^2}{\lambda_{kl}}\right). \tag{3.6}$$

This has the advantage of being an uncertainty model not just of the magnitude but of the complex valued Fourier coefficient, which allows the propagation back into the time domain. It also arises naturally in conventional MMSE speech enhancement methods like the amplitude, log-amplitude or Wiener estimators as seen in [4].

The complex Gaussian uncertainty model can be related to the Gaussian uncertainty model used in [20]. Both models originate, in fact, from a non-central Chi distribution uncertainty model for the amplitude with two and one degrees of freedom, respectively [1]. The approaches to uncertainty propagation introduced in the next chapter will, however, be centered on the complex Gaussian model due the particular advantages previously discussed. A detailed discussion of the methods can be found in [1].

3.3 Piecewise Uncertainty Propagation

The problem of propagating uncertainty corresponds to that of transforming a random variable through a non-linear transformation, in this case the feature extraction. There are many possible methods to compute this transformation, however, there are two characteristics of the particular problem of uncertainty propagation for robust automatic speech recognition that delimit the set of useful techniques:

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.
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, black-box approaches like pseudo-Monte Carlo methods or universal non-linear approximators are not dependent on the complexity of the transformations involved. Such methods suffer, how-

[1] Despite this fact, only the variance was considered for propagation.

ever, from other shortcomings like the need for training with suitable noise samples, or poor trade-offs between accuracy and speed under certain conditions.

Given this context, neither closed-form nor black-box solutions can provide a general solution for the propagation of uncertainty. There exists, however, the possibility of combining both techniques in a piecewise approach that will provide a set of solutions for some well known feature extractions. In this approach, 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 features in the intermediate steps. On the positive side, we can optimize each process individually and share transformations among different feature extractions.

In [1], the piecewise approach to uncertainty propagation was applied to the Mel frequency cepstral and RASTA-LPCC feature extractions, as well as the additional transformations used in the advanced front-end specified in the ES 202 050 standard of the European Telecommunications Standards Institute (ETSI). It was also demonstrated that for the complex Gaussian model, the resulting uncertain features were accurately described by a Gaussian distribution. The next sections will detail the individual solutions used for these transformations as well as the general solutions obtainable for linear transformations and through pseudo-Monte Carlo methods. The last section will present some of the experiments on the accuracy of the piecewise approach and its results in robust automatic speech recognition.

3.3.1 Uncertainty Propagation through Linear Transformations

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

$$E\left\{\mathbf{T}\mathbf{x}^T\right\} = \mathbf{T}E\left\{\mathbf{x}\right\}^T, \tag{3.7}$$

and

$$\mathrm{Cov}\left\{\mathbf{T}\mathbf{x}^T\right\} = \mathbf{T}\mathrm{Cov}\left\{\mathbf{x}\right\}\mathbf{T}^T. \tag{3.8}$$

Furthermore, if a variable is Gaussian distributed, it retains this condition when linearly transformed. However, it has to be taken into account that when $\mathbf{T}$ is not diagonal, the linear transformation induces a non-diagonal covariance, which has to be considered in subsequent transformations.

3.3.2 Uncertainty Propagation with the Unscented Transform

Monte Carlo and pseudo-Monte Carlo 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 [32] is a pseudo-Monte Carlo 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

$$\mathbf{y}_l = \mathbf{g}(\mathbf{x}_l). \tag{3.9}$$

be a non-linear vector valued function. Let $\mathbf{x}_l$ be a frame containing a size N multivariate random variable of distribution $p(\mathbf{x}_l)$ with mean and covariance matrix equal to $\boldsymbol{\mu}_l^x$ and $\boldsymbol{\Sigma}_l^x$, respectively. The result of transforming $\mathbf{x}_l$ through $\mathbf{g}()$ is the multivariate random variable $\mathbf{y}_l$ of distribution $p(\mathbf{y}_l)$. The unscented transform approximates the first and second oder moments of $p(\mathbf{y}_l)$ 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 $p(\mathbf{x}_l)$. The mean and covariance of $\mathbf{y}_l$ can be approximated by using the following weighted averages

$$E\{\mathbf{y}_l\} \approx \sum_{i=1}^{2N+1} W_i \cdot \mathbf{g}(\mathbf{S}_i), \tag{3.10}$$

$$\mathrm{Cov}\{\mathbf{y}_l\} \approx \sum_{i=2}^{2N+1} W_i \cdot (\mathbf{g}(\mathbf{S}_i) - E\{\mathbf{y}_l\}) \cdot (\mathbf{g}(\mathbf{S}_i) - E\{\mathbf{y}_l\})^T. \tag{3.11}$$

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

$$\begin{cases} W_1 = & \frac{\kappa}{(N+\kappa)} \\ \mathbf{S}_1 = & \boldsymbol{\mu}_l^x \\ \mathbf{S}_i = & \boldsymbol{\mu}_l^x + \left(\sqrt{(N+\kappa)\cdot\boldsymbol{\Sigma}_l^x}\right)_i \\ \mathbf{S}_{i+N} = & \boldsymbol{\mu}_l^x - \left(\sqrt{(N+\kappa)\cdot\boldsymbol{\Sigma}_l^x}\right)_i \\ W_i = & \frac{1}{2(N+\kappa)} \quad \text{for} \quad i \in \{2\cdots N+1\}, \end{cases} \tag{3.12}$$

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 $p(\mathbf{x}_l)$ to be modeled. If the option chosen is to assume a Gaussian distribution for $p(\mathbf{x}_l)$, this leads to

$$\kappa = 3 - N. \tag{3.13}$$

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 $p(\mathbf{x}_l)$ can be problematic in the context of uncertainty propagation.

3.4 Uncertainty Propagation for the Mel Frequency Cepstral Features

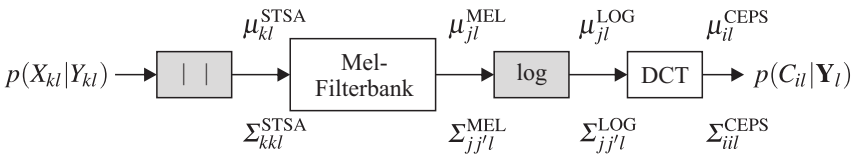

Fig. 3.3: Steps involved in the Mel frequency cepstral feature extraction as well as moments being propagated. Shaded blocks denote non-linear transformations.

The Mel frequency cepstral coefficients (MFCCs) are computed from the STFT of a signal X_{kl} as the following non-linear transformation (see Fig. 3.3)

$$C_{il} = \sum_{j=1}^{J} T_{ij} \log\left(\sum_{k=1}^{K} W_{jk}|X_{kl}|^{\alpha}\right), \tag{3.14}$$

where W_{jk} is the Mel-filterbank weight and

$$T_{ij} = \sqrt{\frac{2}{J}} \cos\left(\frac{\pi i}{J}(j-0.5)\right), \tag{3.15}$$

the discrete cosine transform (DCT) weight. The exponent α usually takes the values 1 or 2, corresponding to the use of the short-time spectral amplitude (STSA) or the estimated power spectral density (PSD) respectively. This transformation is also very similar to the initial steps of the RASTA-PLP transformation later considered. The case of the STSA will be therefore be discussed here, whereas the case of the PSD will be commented on in the RASTA-PLP section.

3.4.1 Propagation through the STSA

If we consider the complex Gaussian Fourier uncertainty model introduced in Section 3.1, we have that each Fourier coefficient of the clean signal X_{kl} is circularly symmetric complex Gaussian distributed. The posterior distribution of each Fourier coefficient given the available information is thus given by Equation (3.6), where $\hat{X}_{kl}$ is the estimated Fourier coefficient of the clean signal given in Equation (3.2) and λ_{kl} the associated variance or uncertainty of estimation given in Equation (3.3). If we express this uncertain Fourier coefficient in polar coordinates we have

$$X_{kl} = A_{kl}\, e^{j\alpha_{kl}}, \tag{3.16}$$

where j corresponds here to the imaginary unit. The distribution of its amplitude

$$A_{kl} = |X_{kl}|, \tag{3.17}$$

can be the obtained by marginalizing the phase in polar coordinates from Equation (3.6) as

$$p(A_{kl}|Y_{kl}) = \int p(X_{kl}|Y_{kl}) d\alpha_{kl} = \int p(A_{kl}, \alpha_{kl}|Y_{kl}) d\alpha_{kl}, \tag{3.18}$$

which leads to the Rice distribution

$$p(A_{kl}|Y_{kl}) = \frac{2A_{kl}}{\lambda_{kl}} \exp\left(-\frac{A_{kl}^2 + |\hat{X}_{kl}|^2}{\lambda_{kl}}\right) I_0\left(\frac{2A_{kl}|\hat{X}_{kl}|}{\lambda_{kl}}\right), \tag{3.19}$$

where I_0 is the modified Bessel function of order zero obtained by applying [15, Eq. (8.431.5)]. The n^{th} order moment of a Rice distribution has the following closed form solution obtained by applying [15, Eqs. 6.631.1, 8.406.3, 8.972.1]

$$E\{A_{kl}^n|Y_{kl}\} = \Gamma\left(\frac{n}{2}+1\right)(\lambda_{kl})^{\frac{n}{2}}\, \mathrm{L}_{\frac{n}{2}}\left(-\frac{|\hat{X}_{kl}|^2}{\lambda_{kl}}\right). \tag{3.20}$$

where $\Gamma()$ is the Gamma function and $\mathrm{L}_{\frac{n}{2}}$ the Laguerre polynomial [2].

[2] Note that the upper index in the Laguerre polynomial is omitted since it is always zero.

Given a general closed-form solution of the n^{th} order moment, the mean of the uncertain STSA can be computed as

$$\mu_{kl}^{\mathrm{STSA}} = E\{A_{kl}|Y_{kl}\} = \Gamma(1.5)\sqrt{\lambda_{kl}}\mathrm{L}_{\frac{1}{2}}\left(-\frac{|\hat{X}_{kl}|^2}{\lambda_{kl}}\right) \tag{3.21}$$

where $\mathrm{L}_{\frac{1}{2}}()$ can be expressed in terms of Bessel functions by combining [15, Eq. 8.972.1] and [27, Eq. 4B-9] as

$$\mathrm{L}_{\frac{1}{2}}(x) = \exp\left(\frac{x}{2}\right)\left[(1-x)I_0\left(-\frac{x}{2}\right) - xI_1\left(-\frac{x}{2}\right)\right]. \tag{3.22}$$

Note the logical similarity between (3.21) and the MMSE-STSA estimator given in [12], 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, both formulas coincide [4].

The variance of the uncertain STSA can be computed in a similar form, using the relation between second order cumulants and moments [25, Eq. 2.4] as

$$\Sigma_{kkl}^{\mathrm{STSA}} = E\{A_{kl}^2|Y_{kl}\} - E\{A_{kl}|Y_{kl}\}^2, \tag{3.23}$$

where the second order moment can be obtained from Equation (3.20) using [15, Eq. 8.970.4] as

$$E\{A_{kl}^2|Y_{kl}\} = \lambda_{kl} + |\hat{X}_{kl}|^2, \tag{3.24}$$

leading to

$$\Sigma_{kkl}^{\mathrm{STSA}} = \lambda_{kl} + |\hat{X}_{kl}|^2 - \left(\mu_{kl}^{\mathrm{STSA}}\right)^2. \tag{3.25}$$

The covariance of each STSA frame $\mathbf{A}_l$ is diagonal since each Fourier coefficient is considered statistically independent of the others.

3.4.2 Propagation through the Mel-Filterbank

The propagation through the Mel-filterbank transformation is much simpler, since it is a linear transformation. The computation of 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 (3.7) and (3.8) for $\mathbf{T} = \mathbf{W}$ yielding

$$\mu_{jl}^{\mathrm{Mel}} = \sum_{k=1}^{K} W_{jk}\mu_{kl}^{\mathrm{STSA}} \tag{3.26}$$

and

$$\Sigma^{\text{Mel}}_{jj'l} = \sum_{k=1}^{K}\sum_{k'=1}^{K} W_{jk}W_{j'k'}\Sigma^{\text{STSA}}_{kk'l} = \sum_{k=1}^{K} W_{jk}W_{j'k}\Sigma^{\text{STSA}}_{kkl}, \tag{3.27}$$

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

3.4.3 Propagation through the Logarithm of the Mel-STSA

After the Mel-filterbank, the uncertain features of each frame follow an unknown distribution (the sum of Rice random variables of different parameters [1]) 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 [20] by approximating the Mel-STSA uncertain features with the log-Normal distribution. This same distribution had also been originally used in [14] to propagate a model of corrupted speech from spectral to logarithm domain and later in [30] for the propagation of noise variance statistics. However, 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 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[3] favors the use of the pseudo-Monte Carlo method known as the unscented transform, presented in Section 3.3.2.

Indeed, in the particular case of the Mel-STSA uncertain features, the unscented transform can provide a more accurate estimate than the log-Normal assumption and is therefore used, yielding

$$\mu^{\text{LOG}}_{jl} \approx \sum_{i=1}^{2N+1} W_i \cdot \log(S_{ji}) \tag{3.28}$$

$$\Sigma^{\text{LOG}}_{jj'l} \approx \sum_{i=2}^{2N+1} W_i \cdot (\log(S_{ji}) - \mu^{\text{LOG}}_{jl}) \cdot (\log(S_{j'i}) - \mu^{\text{LOG}}_{j'l}). \tag{3.29}$$

where each of the $2N+1$ vector sigma points $\mathbf{S}_i$ was obtained using Equations (3.12) and (3.13) with $\boldsymbol{\mu}^x_l = \boldsymbol{\mu}^{\text{MEL}}_l$, $\boldsymbol{\Sigma}^x_l = \boldsymbol{\Sigma}^{\text{MEL}}_l$ and N is equal to the number of Bark filterbanks J.

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

3.4.4 Propagation through the Discrete Cosine Transform

The cosine transform is a linear transformation and therefore can be easily computed by using Equations (3.7) and (3.8) as

$$\mu_{il}^{\mathrm{CEPS}} = \sum_{j=1}^{J} T_{ij} \mu_{jl}^{\mathrm{LOG}} \tag{3.30}$$

and

$$\Sigma_{iil}^{\mathrm{CEPS}} = \sum_{j=1}^{J} \sum_{j'=1}^{J} T_{ij} T_{ij'} \Sigma_{jj'l}^{\mathrm{LOG}}. \tag{3.31}$$

Note that despite the fact that the full covariance after the DCT transformation can be easily obtained, only the values of the diagonal were used in the experiments to reduce the computational load[4]. Note also that the eventual use of cepstral liftering poses no difficulty, since it is also a linear transformation that can be included as a factor in the transform matrix $\mathbf{T}$.

3.5 Uncertainty Propagation for the RASTA-PLP Cepstral Features

The perceptual linear prediction (PLP) features [16] will first be described without uncertainties for clarity. Subsequent Sections 3.5.1 through 3.5.5 will detail each step of the computation. The feature extraction is depicted in detail in Fig. 3.4, and as it can be seen, it shares the initial steps with the MFCC feature extraction:

$$B_{jl} = \sum_{k=1}^{K} V_{jk} |X_{kl}|^{\alpha}, \tag{3.32}$$

where the V_{kl} correspond to the weights of a Bark filterbank, similar to the Mel filterbank, and α will be equal to 2. The next transformation is the application of an equal loudness pre-emphasis ψ_j and the cubic root compression to obtain a perceptually modified power spectrum

$$H_{jl} = \left(\psi_j B_{jl}\right)^{0.33}, \tag{3.33}$$

where ψ_j is a multiplicative gain in the linear domain that only varies with the filterbank index j. The final step of the RASTA-PLP feature extraction is the computation of the autoregressive all-pole model. Its implementation is done via the inverse DFT and the Levinson-Durbin recursion to obtain the LPCs [8, pp. 290-302].

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

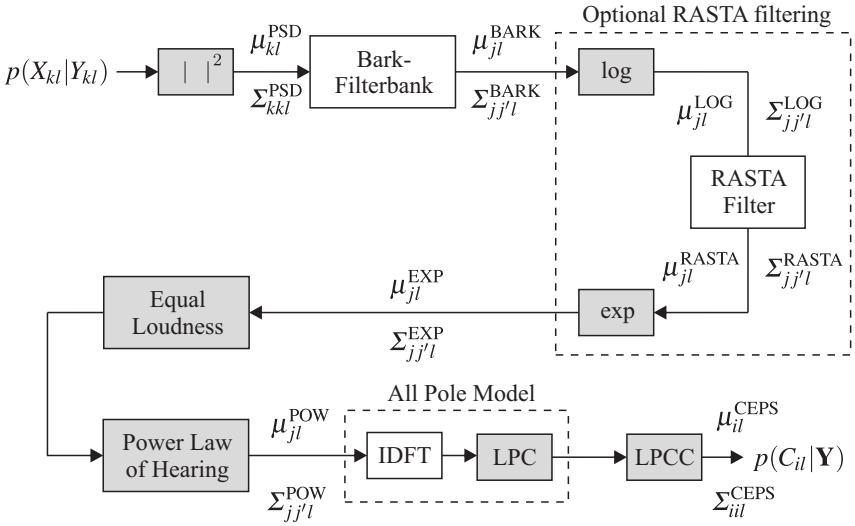

Fig. 3.4: Steps involved in the perceptual linear prediction feature extraction with optional RASTA filtering as well as moments being propagated. Shaded blocks denote non-linear transformations.

As an optional postprocessing step to improve feature performance, LPC based cepstral coefficients (LPCCs) can also be computed using [7, Eq. 3].
A possible proposed improvement of the RASTA-LPCCs is the use the application of a RelAtive SpecTrAl (RASTA) filter to the logarithmic output of the filterbank [16]. This process implies transforming the output of the filterbank through the logarithm and applying an infinite impulse response (IIR) filter that operates over the time dimension (index l) of the feature matrix. This filter corresponds to the difference equation

$$y_{jl} = \sum_{d=1}^{4} b_d x_{jl-d} - a_1 y_{jl-1}, \tag{3.34}$$

where y_{jl} and y_{jl-1} are the l^{th} and $(l-1)^{th}$ RASTA filter outputs and $x_{jl} \cdots x_{jl-4}$ correspond to the four previous logarithm domain inputs of $\log(B_{jl})$. The scalars a_1 and $b_0 \cdots b_4$ are the normalized feedforward and feedback coefficients. After this filtering in the logarithm domain, the signal is transformed back using the exponential transformation; and the PLP transformations is completed.

3.5.1 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 μ_{kl}^{PSD} of the uncertain PSD features can be computed from Equation (3.24) as

$$\mu_{kl}^{\mathrm{PSD}} = E\{A_{kl}^2|Y_{kl}\} = \lambda_{kl} + |\hat{X}_{kl}|^2, \tag{3.35}$$

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

$$\Sigma_{kkl}^{\mathrm{PSD}} = E\{A_{kl}^4|Y_{kl}\} - E\{A_{kl}^2|Y_{kl}\}^2 = 2\lambda_{kl}|\hat{X}_{kl}|^2 + \lambda_{kl}^2, \tag{3.36}$$

where $E\{A_{kl}^4|Y_{kl}\}$ is obtained by solving Equation (3.20) for $n = 4$ using [15, Eq. 8.970.5].

3.5.2 Propagation through the Bark Filterbank

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

$$\mu_{jl}^{\mathrm{BARK}} = \sum_{k=1}^{K} V_{jk}\mu_{kl}^{\mathrm{PSD}} \tag{3.37}$$

and

$$\Sigma_{jj'l}^{\mathrm{BARK}} = \sum_{k=1}^{K}\sum_{k'=1}^{K} V_{jk}V_{j'k'}\Sigma_{kk'l}^{\mathrm{PSD}} = \sum_{k=1}^{K} V_{jk}V_{j'k}\Sigma_{kkl}^{\mathrm{PSD}} \tag{3.38}$$

where V_{jk} was defined in Section 3.5.

3.5.3 Propagation through the Logarithm of the Bark-PSD

From simulation tests, it can be observed that the distribution of the uncertain Bark-PSD features is more skewed than that of the Mel-STSA features. It can also be observed that the log-Normality assumption offers a good approximation for the distribution of the filterbank-PSD features regardless of which of the two filterbanks is used [1]. The log-Normality assumption, also used in other propagation approaches [14, 20, 30], is equivalent to the assumption of the logarithm of the uncertain Bark-PSD features being Gaussian distributed [18, Chapter 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 transforma-

tion and matching first and second moments, thus obtaining [14, Eqs C. 11 and C. 12]

$$\mu_{jl}^{\mathrm{LOG}} = \log(\mu_{jl}^{\mathrm{Bark}}) - \frac{1}{2}\log\left(\frac{\Sigma_{ijl}^{\mathrm{BARK}}}{\mu_{jl}^{\mathrm{BARK}}\mu_{jl}^{\mathrm{BARK}}} + 1\right), \tag{3.39}$$

and

$$\Sigma_{jj'l}^{\mathrm{LOG}} = \log\left(\frac{\Sigma_{jj'l}^{\mathrm{BARK}}}{\mu_{jl}^{\mathrm{Bark}}\mu_{j'l}^{\mathrm{Bark}}} + 1\right). \tag{3.40}$$

3.5.4 Propagation through RASTA Filter

As an IIR filter, the RASTA-filter is a recursive linear transformation corresponding to the difference equation given in Equation (3.34). Computing the propagation of the mean through this transformation is immediate due to the linearity of the expectation operator. Applying Equation (3.7), we obtain

$$\boldsymbol{\mu}_l^{\mathrm{RASTA}} = E\{\mathbf{y}_l\} = \sum_{d=0}^{4} b_d \boldsymbol{\mu}_l^{\mathrm{LOG}} - a_1 \boldsymbol{\mu}_{l-1}^{\mathrm{RASTA}}, \tag{3.41}$$

where the last term $\boldsymbol{\mu}_{l-1}^{\mathrm{RASTA}}$ was obtained in the previous frame computation. The computation of the covariance is more complex, however, 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 [25, Eq. 2.4]

$$\boldsymbol{\Sigma}_l^{\mathrm{RASTA}} = E\{\mathbf{y}_l\mathbf{y}_l^T\} - E\{\mathbf{y}_l\}E\{\mathbf{y}_l\}^T, \tag{3.42}$$

where the second term can be directly calculated from Equation (3.41) but the first term must be solved. By replacing (3.34) into $E\{\mathbf{y}_l\mathbf{y}_l^T\}$ we obtain

$$E\{\mathbf{y}_l\mathbf{y}_l^T\} = E\left\{\left(\sum_{d=0}^{4} b_d\mathbf{x}_{l-d} - a_1\mathbf{y}_{l-1}\right)\left(\sum_{d=0}^{4} b_d\mathbf{x}_{l-d} - a_1\mathbf{y}_{l-1}\right)^T\right\}. \tag{3.43}$$

Taking into account the linearity of the expectation operator, the distributivity of matrix multiplication, and the properties of the transpose operator, we can further simplify this as

$$E\{\mathbf{y}_l\mathbf{y}_l^T\} = \sum_{d=0}^{4} b_d^2 E\{\mathbf{x}_{l-d}\mathbf{x}_{l-d}^T\} + a_1^2 E\{\mathbf{y}_{l-1}\mathbf{y}_{l-1}^T\} - \sum_{d=0}^{4} b_d a_1 E\{\mathbf{y}_{l-1}\mathbf{x}_{l-d}^T\} - \sum_{d=0}^{4} b_d a_1 E\{\mathbf{x}_{l-d}\mathbf{y}_{l-1}^T\}, \tag{3.44}$$

where the last two terms are the transpose of each other and account for the correlation between the 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, for example, we take the term $E\{\mathbf{y}_{l-1}\mathbf{x}_{l-d}^T\}$, we can further expand it by inserting Equation (3.34) shifted by one time unit, which leads to

$$\begin{aligned} E\{\mathbf{x}_{l-d}\mathbf{y}_{l-1}^T\} &= E\left\{\mathbf{x}_{l-d}\left(\sum_{p=0}^{4} b_p \mathbf{x}_{l-1-p} - a_1\mathbf{y}_{l-2}\right)^T\right\} \\ &= \sum_{p=0}^{4} b_p E\{\mathbf{x}_{l-d}\mathbf{x}_{l-1-p}^T\} - a_1 E\{\mathbf{x}_{l-d}\mathbf{y}_{l-2}^T\}. \end{aligned} \tag{3.45}$$

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

$$E\{\mathbf{x}_{l-d}\mathbf{y}_{l-1}^T\} = b_{d-1} E\{\mathbf{x}_{l-d}\mathbf{x}_{l-d}^T\} - a_1 E\{\mathbf{x}_{l-d}\mathbf{y}_{l-2}^T\}. \tag{3.46}$$

The second term of this formula is again an input and output correlation that can be expanded as in Equation (3.45) and simplified the same way. By recursively replacing the time shifted equivalents of Equation (3.34) into the remaining input output correlation we reach a point, at which the shifted output is no longer dependent on the inputs, thus obtaining

$$\begin{aligned} E\{\mathbf{x}_{l-d}\mathbf{y}_{l-1}^T\} &= b_{d-1} E\{\mathbf{x}_{l-d}\mathbf{x}_{l-d}^T\} - a_1 b_{d-2} E\{\mathbf{x}_{l-d}\mathbf{x}_{l-d}^T\} + a_1^2 E\{\mathbf{x}_{l-d}\mathbf{y}_{l-3}^T\} \\ &\quad + \sum_{q=1}^{d} (-1)^{q-1} a_1^{q-1} b_{d-q} E\{\mathbf{x}_{l-d}\mathbf{x}_{l-d}^T\}. \end{aligned} \tag{3.47}$$

Substituting this result into Equation (3.44) and replacing this last equation as well as the propagated mean in Equation (3.41) into Equation (3.42) we obtain the following formula for the covariance

$$\boldsymbol{\Sigma}_l^{\text{RASTA}} = \sum_{d=0}^{4} b_d^2 E\{\mathbf{x}_{l-d}\mathbf{x}_{l-d}^T\} + a_1^2 E\{\mathbf{y}_{l-1}\mathbf{y}_{l-1}^T\} + 2\sum_{d=0}^{4} b_d \sum_{q=1}^{d} (-1)^q a_1^q b_{d-q} E\{\mathbf{x}_{l-d}\mathbf{x}_{l-d}^T\} - \boldsymbol{\mu}_l^{\text{RASTA}} \left(\boldsymbol{\mu}_l^{\text{RASTA}}\right)^T, \quad (3.48)$$

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 always be solved using the conventional matrix solutions in Equations (3.7) and (3.8), 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.

3.5.5 Propagation through Equal Loudness Pre-Emphasis, Power Law of Hearing and Exponential

Both equal loudness pre-emphasis and power law of hearing transformations, given in (3.33), can be performed in the log domain where they can be formulated as the following linear transformation

$$H_{jl} = \exp(0.33\log(\psi_j) + 0.33\log(B_{jl})). \quad (3.49)$$

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 equal loudness 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, in Equations (3.7) and (3.8), with the log-Normal forward propagation through the exponential given in [14, Eqs C. 5 and C. 8]

$$\mu_{jl}^{\text{EXP}} = \exp\left(0.33\log(\psi_j) + 0.33\mu_{jl}^{\text{RASTA}} + \frac{0.33^2\Sigma_{jjl}^{\text{RASTA}}}{2}\right), \quad (3.50)$$

and

$$\Sigma^{\mathrm{EXP}}_{jj'l} = \mu^{\mathrm{EXP}}_{jl}\mu^{\mathrm{EXP}}_{j'l}\left(\exp\left(0.33^2\Sigma^{\mathrm{RASTA}}_{jj'l}\right)-1\right). \tag{3.51}$$

3.5.6 Propagation through the All Pole Model and LPCC Computation

As discussed at the beginning of this section, 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 of these transformations in one single non-linear transformation $\mathbf{g}$. This is advantageous 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 mean and covariance can be performed using the corresponding unscented transform Equations (3.10) and (3.11) with $\boldsymbol{\mu}^x_l = \boldsymbol{\mu}^{\mathrm{EXP}}_l$, and $\boldsymbol{\Sigma}^x_l = \boldsymbol{\Sigma}^{\mathrm{EXP}}_l$

3.6 Uncertainty Propagation for the ETSI ES 202 050 Standard

As an example of a state-of-the-art robust feature extraction, the propagation problem was solved for the advanced front-end of the ETSI standard. 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 3.5 and thus the uncertainty propagation approach explained in this section remains valid when using Equations (3.35), (3.36), (3.37), (3.38), (3.39), (3.40), (3.30) and (3.31) and taking into account the change between Bark and Mel matrices. In addition to these already introduced transformations, the advanced front-end introduces two additional transformations: the log-energy and a blind equalization step.

The log-energy is used as a measure of energy obtained directly from the time domain raw data. This is computed from each time domain frame of the STFT (see Equation (3.1))

$$E_l = \log\left(\sum_{t'=1}^{N} x\left(t' + (l-1)M\right)^2\right), \tag{3.52}$$

where $t' \in [1 \cdots N]$ is the sample index within the frame and M the frame shift.

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 [13, Section 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

$$\tilde{C}_{il} = C_{il} - b_{il} \tag{3.53}$$

where b_{il} is a bias correction computed recursively from[5]

$$b_{il} = (1-\mu_l)b_{il-1} + \mu_l \left(C_{il-1} - C_i^{\text{ref}}\right) \tag{3.54}$$

and where the step size factor μ_l is obtained online using the previous frame energy information by applying the following rule

$$\mu_l = \begin{cases} \nu & \text{if } \frac{E_{l-1}-211}{64} >= 1 \\ \nu \cdot \frac{E_{l-1}-211}{64} & \text{if } 0 < \frac{E_{l-1}-211}{64} < 1 \\ 0 & \text{if } \frac{E_{l-1}-211}{64} =< 0, \end{cases} \tag{3.55}$$

with $\nu = 0.0087890625$ being a fixed parameter given in [22].

3.6.1 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 (3.6), it is possible to transform $\mathbf{X}$ 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 $X_{1l} \ldots X_{Kl}$, the corresponding time-domain frame $x_{1l} \ldots x_{Nl}$, where K denotes the number of bins up to half the sampling frequency and $N = 2K-2$ the size of the frame[6]. In order to apply the IDFT, it is first necessary to obtain the whole spectrum (N bins), including the frequency bins at half the sampling frequency. This can be obtained as

$$\mathbf{X}'_l = [X_{1l} \ldots X_{Kl}, X^*_{K-1l} \ldots X^*_{2l}] \tag{3.56}$$

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

[5] Note that the ETSI 202 050 Manual and the C-Code implementation differ in the multiplying factor of the first summand of Equation (3.54). Here, the implementation of the C-Code was used as the reference.

[6] 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 omitted here for clarity.

$$E\{X_{kl}^*\} = E\{X_{kl}\}^*, \tag{3.57}$$

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

$$E\{\mathbf{X}_l'\} = \hat{\mathbf{X}}_l' = [\hat{X}_{1l} \dots \hat{X}_{Kl}, \hat{X}_{K-1l}^* \dots \hat{X}_{2l}^*]. \tag{3.58}$$

The correlations between real and imaginary components of the complex valued coefficients can be obtained using the definition of complex Gaussian distribution (see Equation (3.6)). Considering the real X_{kl}^R and imaginary X_{kl}^I components of the spectrum separately, the resulting diagonals of their correlation matrices yield

$$E\{X_{kl}'^R X_{kl}'^R\} = \mathrm{Var}\{X_{kl}'^R\} + E\{X_{kl}'^R\}^2 = \lambda_{kl}/2 + \left(\hat{X}_{kl}^R\right)^2, \tag{3.59}$$

and

$$E\{X_{kl}'^I X_{kl}'^I\} = \mathrm{Var}\{X_{kl}'^I\} + E\{X_{kl}'^I\}^2 = \lambda_{kl}/2 + \left(\hat{X}_{kl}^I\right)^2, \tag{3.60}$$

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

$$E\{X_{kl}'^R X_{k'l}'^R\} = \lambda_{kl}/2 + \left(\hat{X}_{kl}^R\right)^2, \tag{3.61}$$

and

$$E\{X_{kl}'^I X_{k'l}'^I\} = -\lambda_{kl}/2 - \left(\hat{X}_{kl}^I\right)^2, \tag{3.62}$$

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

$$x_{t'l} = \mathrm{Re}\left\{ \frac{1}{N} \sum_{k=1}^{N} X_{kl}' \exp\left(2\pi j \frac{(t'-1)(k-1)}{N} \right) \right\}, \tag{3.63}$$

where $t' \in [1 \cdots N]$ is the sample index within the frame.

This transformation can be expressed as the following matrix operation[7]

$$\mathbf{x}_l = \mathrm{Re}\left\{ \mathbf{F} \mathbf{X}_l'^T \right\}, \tag{3.64}$$

where

$$F_{kt'} = F_{kt'}^R + j F_{kt'}^I = \frac{1}{N} \exp\left(2\pi j \frac{(t'-1)(k-1)}{N} \right), \tag{3.65}$$

is the inverse Fourier matrix expressed in cartesian and polar coordinates, which may also be extended to include the inverse window function if one is used.

Since X_{kl}' can be decomposed into its real and imaginary components

[7] Re{} operates elementwise over the vector.

$$X'_{kl} = X'^{R}_{kl} + iX'^{I}_{kl}, \tag{3.66}$$

Equation (3.64) can be expressed as the following linear transformation involving real valued matrices

$$\mathbf{x}_l = \mathrm{Re}\left\{\mathbf{F}\mathbf{X}_l^T\right\} = \mathbf{F}^R\left(\mathbf{X}_l'^R\right)^T - \mathbf{F}^I\left(\mathbf{X}_l'^I\right)^T \tag{3.67}$$

where $F^R_{kt}, F^I_{kt} \in \mathbb{R}$, and by the definition of complex Gaussian distribution, $\mathbf{X}'^R_l$ and $\mathbf{X}'^I_l$ are multivariate Gaussian distributed with the correlations given by Equations (3.59), (3.60), (3.61) and (3.62). Consequently, $\mathbf{x}_l$ will be multivariate Gaussian distributed with mean

$$E\{\mathbf{x}_l\} = \mathbf{F}^R\left(\hat{\mathbf{X}}_l'^R\right)^T - \mathbf{F}^I\left(\hat{\mathbf{X}}_l'^I\right)^T \tag{3.68}$$

and correlation

$$E\{\mathbf{x}_l\mathbf{x}_l^T\} = \mathbf{F}^R E\left\{\mathbf{X}_l'^R(\mathbf{X}_l'^R)^T\right\}\left(\mathbf{F}^R\right)^T + \mathbf{F}^I E\left\{\mathbf{X}_l'^I(\mathbf{X}_l'^I)^T\right\}\left(\mathbf{F}^I\right)^T, \tag{3.69}$$

where for this last formulation, the statistical independence of real and imaginary components has been used, and the correlation matrices in Equations (3.59), (3.60), (3.61) and (3.62) 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 [6, Eq. 3]).

3.6.2 Propagation through Log-Energy Transformation

The computation of the log-energy is attained from each time domain frame of the STFT as given by Equation (3.52). As shown in the last section, for the complex Gaussian uncertainty model, each frame of time-domain elements $x_{1l}\cdots x_{Nl}$ is multivariate Gaussian distributed with a covariance matrix given by Equation (3.69). Consequently, the mean of the squared time domain elements $x^2_{1l}\cdots x^2_{Nl}$ will correspond to the diagonal of Equation (3.69). Furthermore, given the formula which relates the fourth order cross cumulant and the raw moments of a random variable [25, Eq. 2.7] and provided that a multivariate Gaussian distribution has a fourth order cross cumulant equal to zero, we have

$$E\{x^2_{t'l}x^2_{t''l}\} = E\{x^2_{t'l}\}E\{x^2_{t''l}\} + 2E\{x_{t'l}x_{t''l}\}^2 - 2E\{x_{t'l}\}^2E\{x_{t''l}\}^2, \tag{3.70}$$

which enables us to compute the full covariance of the squared Gaussian variables from the mean and covariance given by Equations (3.68) and (3.69). To complete the transformation in Equation (3.52), the logarithm of the sum of $x^2_{1l}\cdots x^2_{Nl}$ has to be computed. The sum is a linear transformation and thus can be computed

with Equations (3.7) and (3.8), whereas the logarithm has to be approximated. For this purpose, we are using the log-Normal assumption specified by Equations (3.39) and (3.40). Combining the formulas for both mean and covariance results in

$$\Sigma_l^{\mathrm{LogE}} = \log\left(\sum_{t'=1}^{N}\sum_{t''=1}^{N} E\{x_{t'l}^2 x_{t''l}^2\}\right) - 2\log\left(\sum_{t'=1}^{N} E\{x_{t'l}^2\}\right) \tag{3.71}$$

and

$$\mu_l^{\mathrm{LogE}} = \log\left(\sum_{t'=1}^{N} E\{x_{t'l}^2\}\right) - \frac{1}{2}\Sigma_l^{\mathrm{LogE}}, \tag{3.72}$$

where in this case, for the sake of simplicity, Equation (3.40) has been re-expressed in terms of mean and correlation instead of mean and covariance using [25, Eq. 2.4].

3.6.3 Propagation through Blind Equalization

The propagation of uncertainty through the blind equalization transformation, given by Equations (3.53), (3.54) and (3.55), can be rather complex 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 Equations (3.54) and (3.55)) 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 (3.7) and (3.8) on (3.53) and (3.54) individually. The temporal correlations induced between features of different frames are also ignored here.

3.7 Taking Full Covariances into Consideration

3.7.1 Covariances Between Features of the Same Frame

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 in general.

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 3.4.2, 3.5.2). 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[8]. Furthermore, it also greatly simplifies the unscented transform algorithm (see Section 3.3.2) 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 3.6.2).

The effect of both assumptions will be analyzed in the experimental setup in Section 3.8.

3.7.2 Covariance between Features of Different Frames

Apart from the RASTA filter, there are various linear transformations typically used in feature extractions that combine features of different time-frames. Many such transformations can be solved by applying Equations (3.7) and (3.8), as for example the computations of delta and acceleration features, and cepstral mean subtraction. Nevertheless, if a previous transformation (e.g. a RASTA filter) has induced a correlation between features of different frames, this should be included in the covariance matrix. For the typical output of RASTA filtering, 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 nearby frames, but it remains, nevertheless, a very significant increase in computational cost. Inter-frame correlation is therefore ignored in delta, acceleration and cepstral mean subtraction transformations. As it will be seen in Section 3.8.2, this does not much affect the accuracy of the uncertainty propagation algorithms.

[8] The value of J varies with the implementations but is usually around 20.

3.8 Tests and Results

3.8.1 Monte Carlo Simulation of Uncertainty Propagation

For the assumed complex Gaussian model of uncertainty, given by Equation (3.6), it is very simple to generate random samples of the distribution. This allows for an approximation of the true uncertain feature distributions by using a Monte Carlo simulation. For this purpose a simulation experiment with additive noise at different SNR levels was carried out. For each case, samples of the resulting STFT matrix of complex Gaussian distributed values were drawn and transformed through the different feature extractions. The resulting feature matrices of samples were used to compute different statistics. The kurtosis was used as a measure of Gaussianity to determine average and extreme cases. Two cases were analyzed. Uncertainty propagation considering the full covariance after Mel-filterbank, labeled F, and uncertainty propagation with diagonal covariances after the filterbank, labeled D[9].
Fig. 3.5 shows the uncertain feature distributions representative of the average case. The true distribution of the features obtained through the Monte Carlo simulation (shaded) are compared with the assumed distributions (dashed). The parameters of the assumed distributions are computed from the first and second order moments computed as indicated in Section 3.3 computed by using full covariances. The results show that the hypothesis of Gaussianity of the three explored feature types holds reasonably well in all scenarios.

To complete the analysis, first and second order moments obtained through the Monte Carlo simulation were compared with the ones obtained with the proposed formulas. Regarding the estimation of the mean, the Monte Carlo and proposed propagation estimations were almost always indistinguishable. The accuracy of the mean estimation was also not affected by the use of full (F) or diagonal covariances (D). Regarding the error in the propagated covariance, this was rather low when using full covariance in all scenarios. When using diagonal covariances, the error increased, particularly for certain features, but the estimated variances remained approximately proportional to the true variances. Fig. 3.6 displays the variance of the feature component with the highest absolute variance estimation error, comparing Monte Carlo (gray), full covariance (F, solid black) and diagonal covariance (D, dashed) solutions. The error of variance propagation is most noticeable in the case of RASTA-LPCC features propagated with diagonal covariance (D), although this error is only exhibited by the higher order cepstral coefficients. The use of full covariance (F) provides an accurate computation of the first and second order moments.

[9] See [1] for a detailed explanation of the experiments.

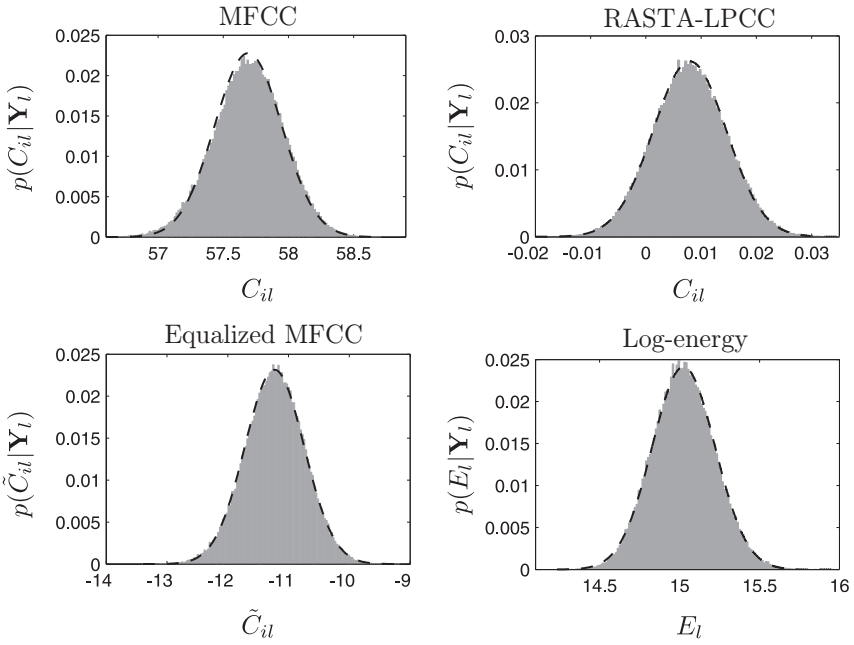

Fig. 3.5: Comparison between uncertainty distributions computed using a Monte Carlo simulation (solid gray) with the assumed uncertainty distributions (dashed) computed from the propagated first and second order moments with full covariance (F). **From left to right, top to bottom:** MFCC, RASTA-LPCC, equalized MFCC according to the ETSI-AFE standard and log-energy.

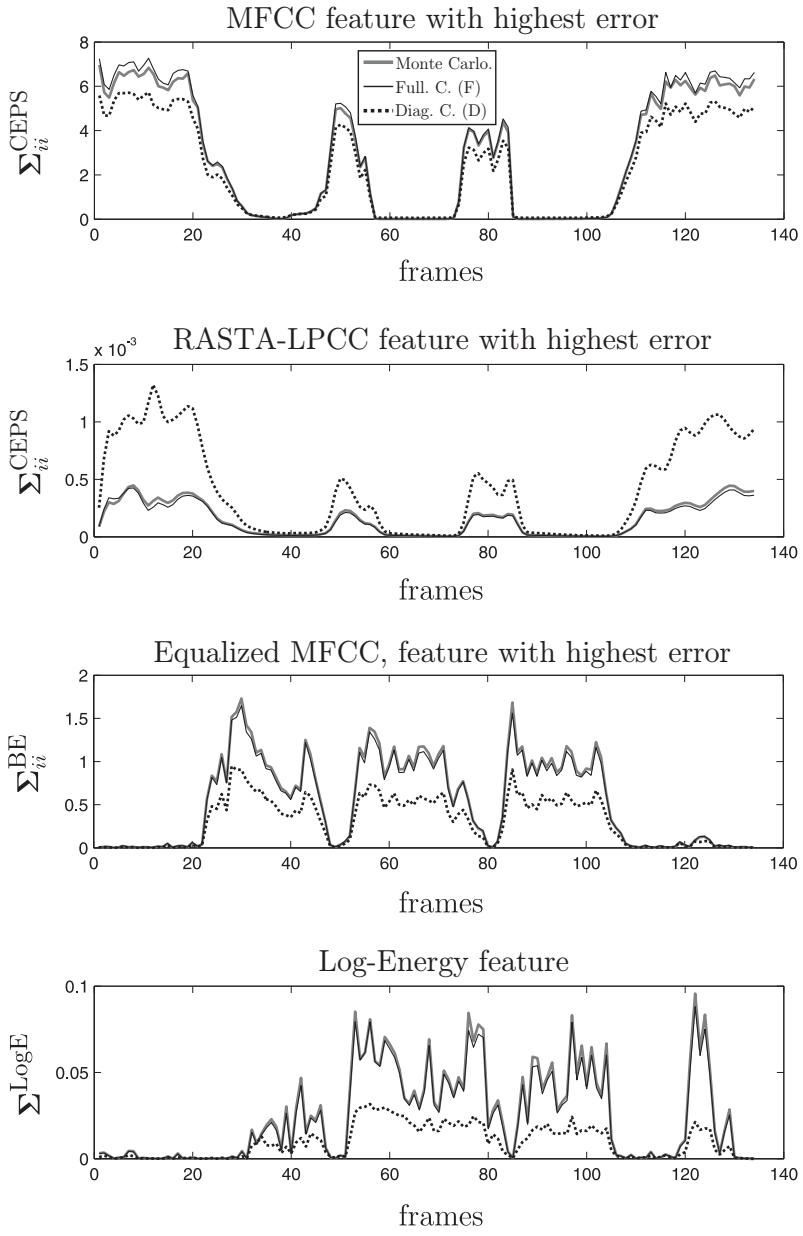

Fig. 3.6: Comparison between propagated variance computed using a Monte Carlo simulation (solid gray) with the covariance computed with the proposed uncertainty propagation algorithms using diagonal (D, dotted black) and full covariance (F, solid black). Only the feature with the highest absolute variance estimation error, when comparing the Monte Carlo simulation with the propagated solution using full covariances, is displayed. **From top to bottom:** MFCC, RASTA-LPCC, equalized MFCC according to the ETSI-AFE standard and log-energy.

3.8.2 Improving the Robustness of ASR with Uncertainty Propagation

The usefulness of the proposed propagation algorithms for robust ASR has been confirmed by positive results in a variety of ASR tasks employing single and multiple channel speech enhancement algorithms in the STFT domain [1, 4, 19]. The results displayed here correspond to the improvement of the European European Telecommunications Standards Institute (ETSI) advanced front-end (AFE) in the AURORA 5 test environment [13], originally described in [2]. The highly optimized AFE combines multiple different noise suppression approaches to achieve very robust features. At certain steps of the algorithm, speech is enhanced through techniques which depend on a good noise suppression to work properly. Since this might not always be the case, the uncertainty is set as proportional to the rate of change those steps inflict.

The generated uncertainty is then propagated using the algorithms presented in the previous sections and employed in the domain of recognition features to achieve a more robust speech recognition. In order to integrate the uncertainty with the ASR system, uncertainty decoding [10], labeled UD, and modified imputation, labeled MI, [20] (see Chapter 12) are used. In addition to this, the two cases considered in the Monte Carlo simulation, diagonal (D) and full (F) covariances are also considered.

The results displayed in Table 3.1 show how the use of uncertainty decoding achieves a general reduction in the word error rates (WER) of the standard's baseline, particularly for the reverberant environments for which the standard is not optimized. Furthermore, the use of diagonal covariances during propagation has only a slightly negative or even a positive effect on the recognition rates. It has to be taken into account, that the computational cost of the uncertainty propagation algorithms for diagonal covariance is close to twice that of the conventional MFCC feature extraction. The computational cost of this step is also very small compared with the previous noise suppression steps. The only noticeable increase in costs of the use of uncertainty propagation is the need for transmitting 28 rather than the 14 features usually computed at the terminal side of the standard.

Table 3.1: 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

References

1. Astudillo, R.F.: Integration of short-time fourier domain speech enhancement and observation uncertainty techniques for robust automatic speech recognition. Ph.D. thesis, Technical Univeristy Berlin (2010)
2. Astudillo, R.F., Kolossa, D., Mandelartz, P., Orglmeister, R.: An uncertainty propagation approach to robust ASR using the ETSI advanced front-end. IEEE JSTSP Special Issue on Speech Processing for Natural Interaction with Intelligent Environments (accepted for publication, 2010)
3. Astudillo, R.F., Kolossa, D., Orglmeister, R.: Propagation of statistical information through non-linear feature extractions for robust speech recognition. In: Proc. MaxEnt2007 (2007)
4. Astudillo, R.F., Kolossa, D., Orglmeister, R.: Accounting for the uncertainty of speech estimates in the complex domain for minimum mean square error speech enhancement. In: Proc. Interspeech (2009)
5. Benítez, M.C., Segura, J.C., Torre, A., Ramírez, J., Rubio, A.: Including uncertainty of speech observations in robust speech recognition. In: Proc. International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 137–140 (2004)
6. Cohen, I., Berdugo, B.: Speech enhancement for non-stationary noise environments. Signal Processing **81**(11), 2403 – 2418 (2001)
7. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust., Speech, Signal Processing **28 (4)**(2), 357– 366 (1980)
8. Deller, J.R., Hansen, J.H.L., Proakis, J.G.: Discrete-Time Processing of Speech Signals. Prentice-Hall, Inc. (1987)
9. Deng, L., Droppo, J., Acero, A.: Exploiting variances in robust feature extraction based on a parametric model of speech distortion. In: Proc. International Conference on Spoken Language Processing (ICSLP) (2002)
10. Droppo, J., Acero, A., Deng, L.: Uncertainty decoding with splice for noise robust speech recognition. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)., vol. 1, pp. I–57–I–60 vol.1 (2002)
11. Ephraim, Y., Cohen, I.: Recent Advancements in Speech Enhancement, pp. 1–22. CRC Press (May 17, 2004)
12. Ephraim, Y., Malah, D.: Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust., Speech, Signal Processing **32**(6), 1109–1121 (1984)
13. 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)
14. Gales, M.J.F.: Model-based technique for noise robust speech recognition. Ph.D. thesis, Gonville and Caius College (1995)
15. Gradshteyn, I.S., Ryzhik, I.: Table of Integrals, Series and Products. Elsevier (2007)
16. Hermansky, H., Morgan, N.: Rasta processing of speech. IEEE Trans. on Speech and Audio Processing **2**(4), 578–589 (1994). DOI 10.1109/89.326616
17. Ion, V., Haeb-Umbach, R.: Improved source modeling and predictive classification for channel robust speech recognition. In: Proc. Interspeech (2006)
18. Johnson, N.L.: Continous Univariate Distributions -1. Wiley Interscience (1970)
19. Kolossa, D., Astudillo, R.F., Hoffmann, E., Orglmeister, R.: Independent component analysis and time-frequency masking for speech recognition in multi-talker conditions. Eurasip Journal on Audio, Speech, and Music Processing (2010)
20. Kolossa, D., Klimas, A., Orglmeister, R.: 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), pp. 82–85 (2005)

21. Kolossa, D., Sawada, H., Astudillo, R.F., Orglmeister, R., Makino, S.: Recognition of convolutive speech mixtures by missing feature techniques for ICA. In: Proc. Asilomar Conference on Signals, Systems, and Computers, pp. 1397–1401 (2006)
22. Kuroiwa, S., Tsuge, S., Ren, F.: 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), pp. 585–590 (2003). DOI 10.1109/NLPKE.2003.1275974
23. Liao, H., Gales, M.: Issues with uncertainty decoding for noise robust automatic speech recognition. Speech Communication **50**(4), 265 – 277 (2008). DOI DOI:10.1016/j.specom.2007.10.004
24. McAulay, R., Malpass, M.: Speech enhancement using a soft-decision noise suppression filter. IEEE Trans. Acoust., Speech, Signal Processing **28**(2), 137–145 (1980)
25. Nikias, C.L., Petropulu, A.P.: Higher-order spectra analysis, a nonlinear signal processing framework. Prentice Hall signal processing series (1993)
26. Raj, B., Stern, R.: Reconstruction of missing features for robust speech recognition. Speech Communication **43**(5), 275–296 (2004)
27. Rice, S.O.: Mathematical Analysis of Random Noise, vol. 23. Bell Telephone Labs Inc. (1944)
28. Srinivasan, S., Wang, D.: A supervised learning approach to uncertainty decoding for robust speech recognition. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. I –I (2006)
29. Srinivasan, S., Wang, D.: Transforming binary uncertainties for robust speech recognition. IEEE Trans. Audio, Speech and Language Processing **15**(7), 2130–2140 (2007)
30. Stouten, V., Van Hamme, H., Wambacq, P.: 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), vol. 1, pp. I–I (2006). DOI 10.1109/ICASSP.2006.1660133
31. Stouten, V., Van hamme, H., Wambacq, W.: Model based feature enhancement with uncertainty decoding for noise robust ASR. Speech Communication. **48(11)**, 1502–1514 (2006)
32. und Uhlmann J.K., J.S.: A general method for approximating nonlinear transformations of probability distributions. Tech. rep., Dept. of Engineering Science, University of Oxford, Oxford, UK (1996)
33. Windmann, S., Haeb-Umbach, R.: Parameter estimation of a state-space model of noise for robust speech recognition. Audio, Speech, and Language Processing, IEEE Transactions on **17**(8), 1577 –1590 (2009)
34. Yoma, N., McInnes, F., Jack, M.: Improving performance of spectral subtraction in speech recognition using a model for additive noise. IEEE Trans. Speech, Audio Processing **6 (6)**, 579–582 (1998)