

Chapter 9

Audio Enhancement and Robustness

Our view (. . .) is that it is an essential characteristic of experimentation that it is carried out with limited resources, and an essential part of the subject of experimental design to ascertain how these should be best applied; or, in particular, to which causes of disturbance care should be given, and which ought to be deliberately ignored.

—Sir Ronald A. Fisher

Once an audio recognition system that functions under idealistic conditions is established, the primary concern shifts towards making it robust in a real-world. The previous chapter touched this issue by illustrating how audio source separation can be exploited to recover a clean speech signal from a mixture. Extraction of the desired signal, however, is not a necessary pre-condition for robust audio recognition. Rather, several options exist for system improvement along the chain of processing, and have proved to be promising especially in the monaural case. Thus, we will next have a look at this issue following the overview given in [1].

First, filtering or spectral subtraction of the signal before can be applied directly after the audio capture. This is realised, for example, in the advanced front-end feature extraction (AFE) or Unsupervised Spectral Subtraction (USS). Then, auditory modelling can be introduced in the feature extraction process. The main influence of noise on audio is irreversible loss of information caused by its random behaviour and a distortion in the feature space that can be compensated by a suited audio representation in the noise condition [2, 3]. Examples of features in this direction include the MFCCs, PLP coefficients [4, 5] or RASTA-PLP features [6, 7] (cf. Sect. 6.2.1). Next along the chain of processing is the option to enhance the extracted features aiming at removal of effects as introduced by noise [8–10]. Exemplary techniques are normalisation methods such as (Cepstral) Mean Subtraction (CMS) [11], MVN [12], or HEQ [9]. Such feature enhancement can also be realised in a model based way, such as by jointly using a Switching Linear Dynamic Model (SLDM) for the dynamic behaviour of audio plus a Linear Dynamic Model (LDM) for additive noise [13]. Later in the chain, one could tailor the learning algorithm to

be able to cope with noisy signal input. Alternatives besides HMMs [14], such as Hidden Conditional Random Fields (HCRF) [15], Switching Autoregressive Hidden Markov Models (SAR-HMMs) [16], or other more general DBN structures provide higher flexibility in modelling. For example, the extension of an SAR-HMM to an Autoregressive Switching Linear Dynamical System (AR-SLDS) [17] allows for an explicit noise model leading to higher noise robustness. Another solution is to match the AM (or even LM) or feature space to noisy conditions. This requires a recogniser trained on noisy audio [18]. However, the performance highly depends on how similar the noise conditions for training and testing are [19]. One can thus distinguish between low to highly matched conditions training. Further, it can be difficult to provide knowledge on the type of noisy condition. This can be eased by so-called multi-condition training, where clean and noisy material with different types of noise is mixed. This is usually not as good as perfectly matched condition between the current test setting and the one learnt previously. However, it provides a good compromise by improving on average over different noise conditions. Besides using noisy material for training, model adaptation can be used to quickly adapt the recogniser to a specific noise condition encountered in the test scenario. This covers widely used techniques such as maximum a posteriori (MAP) estimation [20], maximum likelihood linear regression (MLLR) [21], and minimum classification error linear regression (MCELR) [22].

Given the multiplicity of developed techniques for noise robustness in Intelligent Audio Analysis, a selection of relevant techniques and a good coverage of the different stages along the chain of processing is aimed at in this section. As these techniques are often also tailored to the specific type of noise at hand, relevant special cases such as white noise or babble noise are covered, which are very challenging for speech processing. In the ongoing, let us take a detailed look at the above mentioned options in particular for audio signal preprocessing, feature enhancement, and audio modelling. For the sake of better readability, ‘audio of interest’ such as speech, music, or specific sounds of interest as opposed to noise will partly simply be written as ‘audio’ in this chapter.

9.1 Audio Signal Preprocessing

The preprocessing of the audio signal for its enhancement shall compensate noise influence prior to the feature extraction [23–25]. Apart from explicit BASS as was shown in the last chapter, one of the frequently used audio and particular speech signal preprocessing [26] standards is the advanced front-end feature extraction introduced in [27] based on two-step Wiener filtering in the time domain. Spectral subtraction such as USS [10] can lead to similar effects at lower computational requirements in comparison to Wiener filtering [28, 29]. These techniques can also be subsumed under broader audio signal preprocessing despite being carried out in the (magnitude) spectrogram domain. These two techniques will now be introduced in more detail.

9.1.1 Advanced Front-End Feature Extraction

The processing in the AFE [27] is shown in Fig. 9.1: Subsequent to noise reduction the denoised waveforms are processed and cepstral features are computed and blindly equalised.

Preprocessing in the AFE is based on two-stage Wiener filtering. After denoising in the first stage, a second one carries out additional dynamic noise reduction. In this second stage a gain factorisation unit controls the intensity of filtering dependent on the SNR. Figure 9.2 depicts the components of the two noise reduction cycles: First, a framing takes place. Then, the linear spectrum is estimated per frame, and the power spectral density (PSD) is smoothed along the time axis in the PSD Mean block. An audio activity detection (or VAD in the special case of speech) discriminates between audio and noise, and thus the estimated spectrum of the audio frames and noise are used in the computation of the frequency domain Wiener filter coefficients. To obtain a Mel-warped frequency domain Wiener filter, the linear Wiener filter coefficients are smoothed along the frequency axis using a Mel-filterbank [1]. The Mel-warped Inverse DCT unit (Mel IDCT) determines the impulse response of the Wiener filter prior to the input signal’s filtering. The signal then passes through a second noise reduction cycle using this impulse response. Finally, the DC offset removal block eliminates the constant component of the filtered signal.

The Wiener filter approach in the AFE algorithm has the advantage that noise reduction is carried out on the frame-level. Further, the Wiener filter parameters are adapted to the current SNR. This allows to handle non-stationary noise. Important is, however, an exact audio activity detection (or VAD). This can be particularly demanding in the case of negative SNR levels (cf. e.g., Sect. 10.1.2). Overall, the AFE is a rather complex approach sensible to errors and inaccuracies within the individual estimation and transformation steps [1].

9.1.2 Unsupervised Spectral Subtraction

USS’s [10] spectral subtraction scheme bases on a two-mixture model approach of noisy audio. It aims to distinguish audio and background noise at the magnitude spectrogram level. A probability distribution is used to model audio and noise. For the modelling of background noise on silent parts of the time-frequency plane, one usually assumes white Gaussian behaviour for the real and imaginary parts [30, 31]. This corresponds to a Rayleigh probability density function $f_N(m)$ for noise in the magnitude domain:

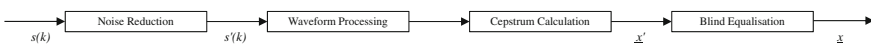


Fig. 9.1 Feature extraction in the AFE according to ETSI ES 202 050 V1.1.5

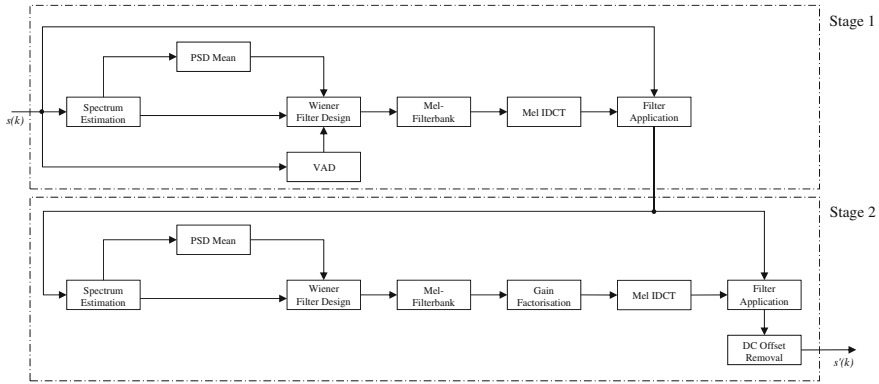


Fig. 9.2 Two-stage Wiener filtering for noise reduction in the AFE according to ETSI ES 202 050 V1.1.5

$$f_N(m) = \frac{m}{\sigma_N^2} e^{-\frac{m^2}{2\sigma_N^2}} \quad (9.1)$$

For the two-mixture model, only an audio ‘activity’ model modelling large magnitudes is needed besides the Rayleigh silence model. For the audio PDF $f_S(m)$ a threshold δ_S is defined with respect to the noise distribution $f_N(m)$ such that only magnitudes $m > \delta_S$ are modelled. In [10], a threshold $\delta_S = \sigma_N$ is used where σ_N is the mode of the Rayleigh PDF. Consequently, magnitudes below σ_N are assumed as background noise. Two additional constraints are needed for $f_S(m)$:

- The derivative $f'_S(m)$ of the ‘activity’ PDF may not be zero if m is just above δ_S ; otherwise the threshold δ_S is meaningless as it could be set to an arbitrarily low value.
- With m towards infinity the decay of $f_S(m)$ should be lower than the decay of the Rayleigh PDF to guarantee $f_S(m)$ modelling large amplitudes.

The ‘shifted Erlang’ PDF with $h = 2$ [32] fulfils these two criteria. It can thus be used to model large amplitudes assumed to be audio of interest:

$$f_S(m) = 1_{m > \sigma_N} \cdot \lambda_S^2 \cdot (m - \sigma_N) \cdot e^{-\lambda_S(m - \sigma_N)} \quad (9.2)$$

with $1_{m > \sigma_N} = 1$ if $m > \sigma_N$ and $1_{m > \sigma_N} = 0$ otherwise.

The overall PDF for the spectral magnitudes of the noisy audio signal is

$$f(m) = P_N \cdot f_N(m) + P_S \cdot f_S(m), \quad (9.3)$$

where P_N is the prior for ‘silence’ and background noise, and P_S is the prior for ‘activity’ and audio of interest. The parameters of the derived PDF $f(m)$ summarised in the set

$$\Lambda = \{P_N, \sigma_N, P_S, \lambda_S\} \quad (9.4)$$

are independent of time and frequency, and can be trained by the EM algorithm (cf. Sect. 7.3.1) [33]. In the expectation step, posteriors are estimated as

$$p(\text{sil}|m_{f,t}, \Lambda) = \frac{P_N \cdot f_N(m_{f,t})}{P_N \cdot f_N(m_{f,t}) + P_S \cdot f_S(m_{f,t})} \quad (9.5)$$

$$p(\text{act}|m_{f,t}, \Lambda) = 1 - p(\text{sil}|m_{f,t}, \Lambda). \quad (9.6)$$

For the Maximisation step, the moment method is used: An update σ_N employing all data takes place before all data with values above the new σ_N help to update λ_S . Two update equations describe the method as follows:

$$\hat{\sigma}_N = \frac{\left[\sum_{f,t} m_{f,t}^2 \cdot p(\text{sil}|m_{f,t}, \Lambda) \right]^{\frac{1}{2}}}{\left[2 \sum_{f,t} p(\text{sil}|m_{f,t}, \Lambda) \right]^{\frac{1}{2}}} \quad (9.7)$$

$$\hat{\lambda}_S = \frac{\sum_{m_{f,t} > \hat{\sigma}_N} (m_{f,t} - \hat{\sigma}_N)^{-1} \cdot p(\text{act}|m_{f,t}, \Lambda)}{\sum_{m_{f,t} > \hat{\sigma}_N} p(\text{act}|m_{f,t}, \Lambda)}. \quad (9.8)$$

Subsequent to the training of all mixture parameters $\Lambda = \{P_N, \sigma_N, P_S, \lambda_S\}$ USS with the parameter σ_N as floor value is applied:

$$m_{f,t}^{USS} = \max \left(1, \frac{m_{f,t}}{\sigma_N} \right) \quad (9.9)$$

Flooring to a non-zero value is required for MFCC or similar features, as zero magnitude values after spectral subtraction can result in unfavourable dynamics. Overall, USS is a simple and efficient preprocessing method that allows for unsupervised EM fitting on observed data. As a downside, it requires reliable estimation of an audio magnitude PDF which is rather challenging. With the PDFs not depending on frequency and time, USS only handles stationary noises. Further, it only models large magnitudes of the audio of interest. Low audio magnitudes thus cannot be distinguished from background noise.

9.2 Feature Enhancement

In feature enhancement, enhancement takes place after the extraction of features to reduce a potential mismatch between test and training conditions. Popular methods include CMS [11], MVN [12], HEQ [9], and the Taylor Series approach [34] able to cope with the non-linear effects of noise. There are some further methods tailored to specific types of features, such as in the cepstrum-domain, where a feature

compensation algorithm to decompose audio of interest and noise is introduced in [35]. To enhance noisy MFCCs, a SLDM can also be used to model the dynamics of audio of interest and those of additive noise by a LDM [13]. An observation model then describes how audio and noise produce the noisy observations to reconstruct the features of clean audio. An extension [36] includes time-dependencies among the discrete state variables of the SLDM. Further, a state model for the dynamics of noise can help to model non-stationary noise sources [37]. Finally, incremental on-line adaptation of the feature space is possible as by feature space maximum likelihood linear regression (FMLLR) [38]. Again, we will now take a detailed look at selected popular approaches.

9.2.1 Feature Normalisation

9.2.1.1 Cepstral Mean Subtraction

To ease the influence of noise and transmission channel transfer functions in cepstral features, CMS [11, 39] provides a simple approach. Its basic principle of mean subtraction can also be applied to practically any other audio LLD. Often, the noise can be considered as comparably stationary when opposed to the rapidly changing characteristics of the audio signal of interest. Thus, a subtraction is carried out of the long-term average cepstral or other feature vector

$$\underline{\mu} = \frac{1}{T} \sum_{t=1}^T \underline{x}_t \quad (9.10)$$

from the observed noise corrupted feature vector sequence of length T :

$$\underline{X} = \{\underline{x}_1, \underline{x}_2, \dots, \underline{x}_t, \dots, \underline{x}_T\} \quad (9.11)$$

By that, a new estimate \tilde{x}_t of the signal in the feature domain results:

$$\tilde{x}_t = \underline{x}_t - \underline{\mu}, \quad 1 \leq t \leq T \quad (9.12)$$

The subtraction of the long-term average is particularly interesting in the cepstral domain. Since the audio spectrum is multiplied by the channel transfer function (cf. Sect. 6.2.1.4), by the logarithm application in the MFCC calculation, this multiplication turns into an addition, and this part can be eliminated by subtraction of the cepstral mean from all input vectors. A disadvantage of CMS, as opposed to HEQ, is the disability to treat non-linear noise effects.

9.2.1.2 Mean and Variance Normalisation

The subtraction of the mean per feature vector component corresponds to an equalisation of the first moment of the vector sequence probability distribution. If noise has also an influence on the variance of the features, according variance normalisation of the vector sequence can be applied and by that an equalisation of the first two moments. This is known as MVN. The processed feature vector is obtained by

$$\tilde{x}_t = \frac{x_t - \mu}{\underline{\sigma}}. \quad (9.13)$$

The division by the vector $\underline{\sigma}$ of the standard deviations per feature vector components is computed out element-by-element. The new feature vector's components have zero mean and unity variance.

9.2.1.3 Histogram Equalisation

HEQ is a popular technique in digital image processing [40] where it helps raise the contrast of images and alleviates the influence of the lighting conditions. In audio processing, HEQ can improve the temporal dynamics of noise-affected feature vector components. HEQ extends the principle of CMS and MVN to all moments of the probability distribution of the feature vector components [9, 41], and by that compensates non-linear distortions caused by noise.

In HEQ, the histogram of each feature vector component is mapped onto a reference histogram. The underlying assumption is that noise influence can be described as a monotonic partly reversible feature transformation. With success depending on meaningful histograms, HEQ requires several frames for their reliable estimation. A key advantage leading to HEQ's independence of the noise characteristics is that no assumptions are made on the statistical properties (e.g., normality) of the noise process.

For HEQ, a transformation

$$\tilde{x} = F(x) \quad (9.14)$$

needs to be found for the conversion of the PDF $p(x)$ of an audio feature into a reference PDF $\tilde{p}(\tilde{x}) = p_{ref}(\tilde{x})$. If x is a unidimensional variable with PDF $p(x)$, a transformation $\tilde{x} = F(x)$ modifies the probability distribution, such that the new distribution of the obtained variable \tilde{x} can be expressed as

$$\tilde{p}(\tilde{x}) = p(G(\tilde{x})) \frac{\partial G(\tilde{x})}{\partial \tilde{x}} \quad (9.15)$$

with $G(\tilde{x})$ as the inverse transformation corresponding to $F(x)$. For the cumulative probabilities based on the PDFs, let us consider:

$$\begin{aligned}
C(x) &= \int_{-\infty}^x p(x') dx' \\
&= \int_{-\infty}^{F(x)} p(G(\tilde{x}')) \frac{\partial G(\tilde{x})}{\partial \tilde{x}'} d\tilde{x}' \\
&= \int_{-\infty}^{F(x)} \tilde{p}(\tilde{x}') d\tilde{x}' \\
&= \tilde{C}(F(x))
\end{aligned} \tag{9.16}$$

By that, the transformation converting the distribution $p(x)$ into the ‘target’ distribution $\tilde{p}(\tilde{x}) = p_{ref}(\tilde{x})$ can be expressed as

$$\tilde{x} = F(x) = \tilde{C}^{-1}[C(x)] = C_{ref}^{-1}[C(x)], \tag{9.17}$$

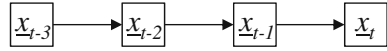
where $C_{ref}^{-1}(\dots)$ is the inverse cumulative probability function of the reference distribution [1]. Further, $C(\dots)$ is the feature’s cumulative probability function. To obtain the transformation per feature vector component, a ‘rule of thumb’ is to use 500 uniform intervals between $\mu_i - 4\sigma_i$ and $\mu_i + 4\sigma_i$ for the derivation of the histograms. μ_i and σ_i are the mean and standard deviation of the i th feature vector element. A Gaussian probability distribution with zero mean and unity variance can be used per element as a reference probability distribution, then, however, ignoring higher moments.

From the feature normalisation strategies discussed above, CMS is the simplest. Together with MVN, it is used most frequently. MVN usually leads to better results at slightly increased computational effort. However, these two techniques both provide a linear transformation. This is different for HEQ, which is able to compensate non-linear effects, but requires sufficient audio frames for good results. HEQ further corrects only monotonic transformations. This can cause an information loss, given that random noise behaviour renders the needed transformation non-monotonic.

9.2.2 Model Based Feature Enhancement

In model based audio enhancement one usually models audio and noise individually plus how these two produce the observation. Then, the features are enhanced to benefit the audio of interest by use of these models. An example is a SLDM to model the dynamics of clean audio of interest [13] that will next be introduced by the mentioned three models for noise, audio, and the combination.

Fig. 9.3 LDM for the modelling of noise



9.2.2.1 Modelling of Noise

Noise is modelled by a simple LDM with the system equation

$$\underline{x}_t = \underline{A} \underline{x}_{t-1} + \underline{b} + \underline{g}_t, \tag{9.18}$$

where the matrix \underline{A} and the vector \underline{b} simulate the noise process’s evolution over time. Further, \underline{g}_t is a Gaussian noise source that drives the system. A graphical model representation of this LDM is given in Fig. 9.3. In this and the following visualisations in this section, squares again indicate observations. With LDMs being time-invariant, they can model signals such as coloured stationary Gaussian noises. The LDM is expressed by

$$p(\underline{x}_t | \underline{x}_{t-1}) = \mathcal{N}(\underline{x}_t; \underline{A} \underline{x}_{t-1} + \underline{b}, \underline{C}) \tag{9.19}$$

$$p(\underline{x}_{1:T}) = p(\underline{x}_1) \prod_{t=2}^T p(\underline{x}_t | \underline{x}_{t-1}), \tag{9.20}$$

where $\mathcal{N}(\underline{x}_t; \underline{A} \underline{x}_{t-1} + \underline{b}, \underline{C})$ is a multivariate Gaussian with the mean vector $\underline{A} \underline{x}_{t-1} + \underline{b}$ and the covariance matrix \underline{C} , and T is the input sequence’s length.

9.2.2.2 Modelling of Audio of Interest

The SLDM models the audio signal of interest passing through states as in a HMM. It further enforces a continuous state transition in the feature space conditioned on the state sequence. This more complex dynamic model has a hidden state variable s_t at each time t . Like this, \underline{A} and \underline{b} depend on the state variable s_t :

$$\underline{x}_t = \underline{A}(s_t) \underline{x}_{t-1} + \underline{b}(s_t) + \underline{g}_t. \tag{9.21}$$

Likewise, the possible state sequences $s_{1:T}$ describe a non-stationary LDM, as \underline{A} and \underline{b} change with time as do the audio features. In Fig. 9.4 the SLDM is shown as graphical model. As one sees, time dependencies are assumed between the continuous

Fig. 9.4 SLDM for the modelling of audio of interest

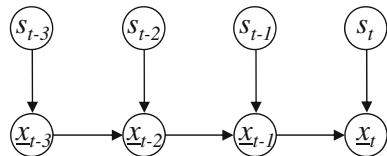
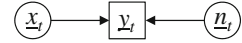


Fig. 9.5 Observation model for noisy audio



variables \underline{x}_t , but not between the discrete state variables s_t [13]. An extension in [36] includes time dependencies between the hidden state variables, similar as in enhancing a GMM to a HMM. A SLDM as in Fig. 9.4 is described by

$$p(\underline{x}_t, s_t | \underline{x}_{t-1}) = \mathcal{N}(\underline{x}_t; \underline{A}(s_t)\underline{x}_{t-1} + \underline{b}(s_t), \underline{C}(s_t)) \cdot p(s_t) \quad (9.22)$$

$$p(\underline{x}_{1:T}, s_{1:T}) = p(\underline{x}_1, s_1) \prod_{t=2}^T p(\underline{x}_t, s_t | \underline{x}_{t-1}). \quad (9.23)$$

The EM algorithm can be used for the learning of the parameters of the SLDM, namely $\underline{A}(s)$, $\underline{b}(s)$, and $\underline{C}(s)$. If one sets the number of states to one the SLDM turns into a LDM to compute the parameters \underline{A} , \underline{b} , and \underline{C} required for the noise modelling LDM.

9.2.2.3 Observation Model

The observation model describes the relationship of the noisy observation \underline{y}_t and the hidden audio and noise features. In Fig. 9.5, the graphical model representation of such a model is given by the zero variance observation model with SNR inference as in [42]. It is assumed that audio of interest \underline{x}_t and noise \underline{n}_t mix linearly in the time domain. In the cepstral domain, for example, this corresponds to a non-linear mixing.

9.2.2.4 Posterior Estimation and Enhancement

To reduce the computational complexity of the posterior estimation, an approximation is given by the restriction of the search space size by the generalised pseudo-Bayesian (GPB) algorithm [43]. It neglects distinct state histories with differences more than r frames in the past. Thus, with T as the sequence length, the inference complexity reduces from S^T to S^r where $r \ll T$. In the GPB algorithm, one ‘collapses’, ‘predicts’, and ‘observes’ for each of the audio frames. Estimates of the moments of \underline{x}_t representing the de-noised audio features are computed based on the Gaussian posterior as calculated during the ‘observation’ in the GPB algorithm. In this process, clean features are assumed to be the Minimum Mean Square Error (MMSE) estimate $E[\underline{x}_t | \underline{y}_{1:t}]$. SLDM feature enhancement can lead to outstanding results including the case of coloured Gaussian noise and negative SNR. This comes by the effort of modelling noise. The audio model’s linear dynamics model the smooth time evolution of typical audio of interest such as speech, music, or certain sound types. The switching states express the piecewise stationarity typical in

such audio. However, noise frames are assumed to be independent over time. As a consequence, non-stationary noises are not modelled adequately. Even with the restrictions made in the GPB algorithm, feature enhancement by SLDM is computationally more demanding than the techniques discussed above. Further, as in the AFE (cf. Sect. 9.1), accurate audio activity detection is required to provide correct estimation of the noise LDM.

9.3 Model Architectures

The most frequently used data-driven model representation of audio are HMMs [14]. Beyond the so far described optimisation options along the chain of Intelligent Audio Analysis, extending HMM topologies to more general DBN layouts can also help to increase noise robustness [15, 17, 44]. Generative models such as HMMs assume conditional independence of the audio feature observations, thus ignoring long-range dependencies as given in most audio of interest [45]. To overcome this, Conditional Random Fields (CRF) [46–48] model a sequence by an exponential distribution given the observation sequence. The HCRF [15, 49] further includes hidden state sequences for the estimation of the conditional probability of a class over an entire sequence. Another interesting option is to model the raw audio signal in the time domain [16]. For example, SAR-HMM [16] provide good results in clean audio conditions. To cope with noise, these can be extended to a Switching Linear Dynamical System (SLDS) [17] to model the dynamics of the raw audio signal and the noise. These alternatives will now be shortly presented.

9.3.1 *Conditional Random Fields*

As mentioned above, CRF [46] use an exponential distribution to model a sequence given its observation and by that also non-local dependencies among states and observations. Further, unnormalised transition probabilities are possible. Owing to the ability to enforce a Markov assumption as in HMMs, dynamic programming is applicable for inference. CRFs were also shown beneficial as LM [50].

9.3.2 *Hidden Conditional Random Fields*

An extension to HCRF is needed to make the CRF paradigm suited for general audio recognition tasks. This comes, as CRF provide a class prediction per observation and frame of a time sequence rather than for an entire sequence. HCRF overcome this by adding hidden state sequences [49]. Reports of superiority over HMM in the Intelligent Audio Analysis domain include the recognition of phones [15] and

non-linguistic vocalisations [51] or the segmentation of meeting speech [52]. A particular strength is the possibility to use arbitrary functions for the observations without complication of the parameter learning.

The HCRF models the conditional probability of a class c , given the sequence of observations $\underline{X} = \underline{x}_1, \underline{x}_2, \dots, \underline{x}_T$:

$$p(c|\underline{X}, \underline{\lambda}) = \frac{1}{z(\underline{X}, \underline{\lambda})} \sum_{Seq \in c} e^{\underline{\lambda} \cdot \underline{f}(c, Seq, \underline{X})}, \quad (9.24)$$

where $\underline{\lambda}$ is the parameter vector and \underline{f} the ‘vector of sufficient statistics’, and $Seq = s_1, s_2, \dots, s_T$ is the hidden state sequence run through during the computation of this conditional probability. The probability is normalised by the ‘partition function’ $z(\underline{X}, \underline{\lambda})$ to ensure a properly normalised probability [15]:

$$z(\underline{X}, \underline{\lambda}) = \sum_c \sum_{Seq \in c} e^{\underline{\lambda} \cdot \underline{f}(c, Seq, \underline{X})}. \quad (9.25)$$

The vector \underline{f} determines the probability to model. With a suited \underline{f} a left-right HMM can be imitated [15]. Let us now restrict the HCRF to a Markov chain, but without the requirements of the transition probabilities to sum to one and real probability densities for the observations. In analogy to a HMM a parametrisation by transition scores $a_{i,j}$ and observation scores $b_j(\underline{x}_t)$ can then be reached with the parameters $\underline{\lambda}$, where i and j are states of the model (cf. Sect. 7.3.2). Forward and backward recursions (cf. Sect. 7.3.1) as for a HMM can then further be used.

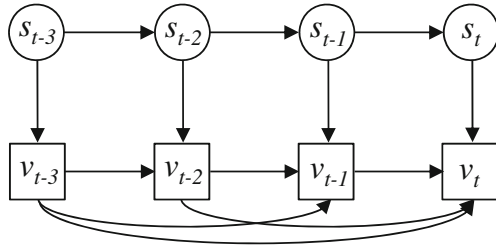
9.3.3 Audio Modelling in the Time Domain

Modelling of the raw signal in the time domain is a sparsely pursued option, but can offer easy explicit noise modelling [16]. We will look at SAR-HMMs to this end first, and then at the extension to SLDS.

9.3.3.1 Switching Autoregressive Hidden Markov Models

The SAR-HMM models the audio signal of interest as an autoregressive (AR) process. The non-stationarity is realised by switching between different AR parameter sets [17] by a discrete switch variable s_t similar to the HMM states. At a time step t —referring to the sample-level in this case—, exactly one out of S states is occupied. The state at time step t depends exclusively on its predecessor with the transition probability $p(s_t|s_{t-1})$. The sample v_t at this time step is assumed as a linear combination of its R preceding samples superposed by a Gaussian distributed ‘innovation’ $\eta(s_t)$. $\eta(s_t)$ and the AR weights $c_r(s_t)$ are the parameter set given by the state s_t :

Fig. 9.6 SAR-HMM as DBN structure



$$v_t = - \sum_{r=1}^R c_r(s_t)v_{t-r} + \eta(s_t) \quad \text{with} \quad \eta \sim \mathcal{N}(\eta; 0, \sigma^2(s_t)). \quad (9.26)$$

There, $\eta(s_t)$ models variations from pure autoregression rather than an independent additive noise process. The joint probability of a sequence of length T is

$$p(s_{1:T}, v_{1:T}) = p(v_1|s_1)p(s_1) \prod_{t=2}^T p(v_t|v_{t-R:t-1}, s_t)p(s_t|s_{t-1}). \quad (9.27)$$

Figure 9.6 visualises the SAR-HMM as DBN structure. Switching of the different AR models is forcedly ‘slowed down’ by introducing an constant K . The model then needs to remain in a state for an integer multiple of time steps. This is needed, as considerably more sample values usually exist than features on the frame level.

The EM algorithm can be used for learning of the AR parameters. Based on the forward-backward algorithm (cf. Sect. 7.3.1) the distributions $p(s_t|v_{1:T})$ are learnt. The fact that an observation v_t depends on R predecessors makes the backward pass more complicated than in the case of an HMM. A ‘correction smoother’ [53] can thus be applied such that the backward pass calculates the posterior $p(s_t|v_{1:T})$ by ‘correcting’ the forward pass’s output.

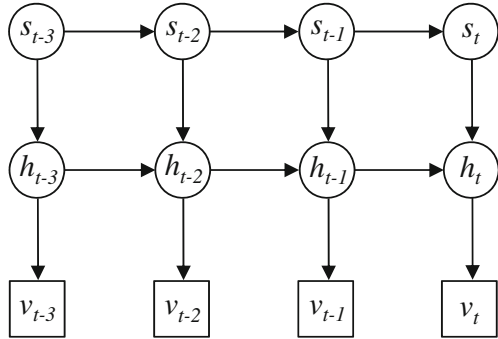
9.3.3.2 Autoregressive Switching Linear Dynamical Systems

With the extension of the SAR-HMM to an AR-SLDS, a noise process can explicitly be modelled [17]. The observed audio sample v_t of interest is then modelled as a noisy version of a hidden clean sample that is obtained from the projection of a hidden vector \underline{h}_t with the dynamic properties of a LDS:

$$\underline{h}_t = \underline{A}(s_t)\underline{h}_{t-1} + \underline{\eta}_t^{\mathcal{H}}, \quad \text{with} \quad \underline{\eta}_t^{\mathcal{H}} \sim \mathcal{N}(\underline{\eta}_t^{\mathcal{H}}; 0, \underline{\Sigma}_{\mathcal{H}}(s_t)). \quad (9.28)$$

The transition matrix $\underline{A}(s_t)$ describes the dynamics of the hidden variable that depends on the state s_t at time step t . A Gaussian distributed hidden ‘innovation’ variable $\underline{\eta}_t^{\mathcal{H}}$ models variations from ‘pure’ linear state dynamics. As for η_t in Eq. (9.26)

Fig. 9.7 AR-SLDS as DBN structure



in the case of the SAR-HMM, $\eta_t^{\mathcal{H}}$ is not modelling an independent additive noise source. For the determination of the observed sample at time step t , the vector \underline{h}_t is projected onto a scalar v_t :

$$v_t = \underline{B} \underline{h}_t + \eta_t^{\mathcal{V}}, \quad \text{with} \quad \eta_t^{\mathcal{V}} \sim \mathcal{N}(\eta_t^{\mathcal{V}}; 0, \sigma_{\mathcal{V}}^2), \quad (9.29)$$

where $\eta_t^{\mathcal{V}}$ models independent additive white Gaussian noise (AWGN) assumed to modify the hidden clean sample $\underline{B} \underline{h}_t$. The DBN structure of the SLDS that models the hidden clean signal and an independent additive noise is found in Fig. 9.7.

The parameters $\underline{A}(s_t)$, \underline{B} and $\underline{\Sigma}_{\mathcal{H}}(s_t)$ of the SLDS can be chosen to mimic the SAR-HMM (cf. Sect. 9.3.3.1) for the case $\sigma_{\mathcal{V}} = 0$ [17]. Likewise, if $\sigma_{\mathcal{V}} \neq 0$ a noise model is included but no training of a new model is needed. With determination of the exact parameters of the AR-SLDS having a complexity of $\mathcal{O}(S^T)$, the Expectation Correction (EC) approximation [54] provides an elegant reduction to $\mathcal{O}(T)$.

In practice, the AR-SLDS is particularly suited to cope with white noise disturbance, as the variable $\eta_t^{\mathcal{V}}$ incorporates an AWGN model. It is, however, usually inferior to frame-level feature-based HMM approaches in clean conditions. This may be explained by the difference of the approach to human perception which is not performed in the time-domain. In coloured noise environment the AR-SLDS usually also leads to lower performance than frame-level feature modelling as by SLDMs. A limitation for practical use is the high computational requirement, even with the EC algorithm: As an example, for audio at 16 kHz, T is 160 times higher than for a feature vector sequence operated on 100 FPS.

Obviously, further model architectures exist that were not shown here, but are well suited to cope with noises, in particular also for non-stationary noise. An example are the LSTM networks as shown in Sect. 7.2.3.4 [55, 56].

References

1. Schuller, B., Wöllmer, M., Moosmayr, T., Rigoll, G.: Recognition of noisy speech: a comparative survey of robust model architecture and feature enhancement. *EURASIP J. Audio Speech Music Process.* (Article ID 942617), 17 (2009)
2. de la Torre, A., Fohr, D., Haton, J.: Compensation of noise effects for robust speech recognition in car environments. In: *Proceedings of International Conference on Spoken Language Processing* (2000)
3. Moreno, P.: Speech recognition in noisy environments. Ph.D. thesis, Carnegie Mellon University, Pittsburgh (1996)
4. Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech. *J. Acoust. Soc. Am.* **87**, 1738–1752 (1990)
5. Junqua, J., Wakita, H., Hermansky, H.: Evaluation and optimization of perceptually-based ASR front-end. *IEEE Trans. Speech Audio Process.* **1**, 329–338 (1993)
6. Hermansky, H., Morgan, N., Bayya, A., Kohn, P.: RASTA-PLP speech analysis technique. In: *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 121–124 (1992)
7. Kingsbury, B., Morgan, N., Greenberg, S.: Robust speech recognition using the modulation spectrogram. *Speech Commun.* **25**, 117–132 (1998)
8. Kim, N.: Nonstationary environment compensation based on sequential estimation. *IEEE Signal Process. Lett.* **5**, 57–59 (1998)
9. de la Torre, A., Peinado, A.M., Segura, J.C., Perez-Cordoba, J.L., Benitez, M.C., Rubio, A.J.: Histogram equalization of speech representation for robust speech recognition. *IEEE Trans. Speech Audio Process.* **13**(3), 355–366 (2005)
10. Lathoud, G., Magimia-Doss, M., Mesot, B., Boulard, H.: Unsupervised spectral subtraction for noise-robust ASR. In: *Proceedings of Automatic Speech Recognition and Understanding*, pp. 189–194 (2005)
11. Rahim, M., Juang, B., Chou, W., Buhcke, E.: Signal conditioning techniques for robust speech recognition. In: *Proceedings of IEEE Signal Processing Letters*, vol. 3, pp. 107–109 (1996)
12. Viikki, O., Laurila, K.: Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Commun.* **25**, 133–147 (1998)
13. Droppo, J., Acero, A.: Noise robust speech recognition with a switching linear dynamic model. In: *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 953–956 (2004)
14. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. In: *Proceedings of the IEEE*, vol. 77, pp. 257–286 (1989)
15. Gunawardana, A., Mahajan, M., Acero, A., Platt, J.C.: Hidden conditional random fields for phone classification. In: *Proceedings of Interspeech*, pp. 1117–1120 (2005)
16. Ephraim, Y., Roberts, W.: Revisiting autoregressive hidden Markov modeling of speech signals. In: *IEEE Signal Processing Letters*, vol. 12, pp. 166–169 (2005)
17. Mesot, B., Barber, D.: Switching linear dynamical systems for noise robust speech recognition. *IEEE Trans. Audio Speech Lang. Process.* **15**, 1850–1858 (2007)
18. Sankar, A., Stolcke, A., Chung, T., Neumeyer, L., Weintraub, M., Franco, H., Beaufays, F.: Noise-resistant feature extraction and model training for robust speech recognition. In: *Proceedings of the 1996 DARPA CSR Workshop* (1996)
19. Macho, D., Mauray, L., Noe, B., Cheng, Y., Ealey, D., Jouviet, D., Kelleher, H., Pearce, D., Saadoun, F.: Evaluation of a noise-robust DSR front-end on Aurora databases. In: *Proceedings of the International Conference on Spoken Language Processing*, pp. 17–20 (2002)
20. Gauvain, J., Lee, C.: Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans. Speech Audio Process.* **2**, 291–298 (1994)
21. Wang, Z., Schultz, T., Waibel, A.: Comparison of acoustic model adaptation techniques on non-native speech. In: *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 540–543 (2003)

22. He, X., Chou, W.: Minimum classification error linear regression for acoustic model adaptation of continuous density HMMs. In: Proceedings of International Conference on Multimedia and Expo, vol. 1, pp. 397–400 (2003)
23. Szymanski, L., Bouchard, M.: Comb filter decomposition for robust ASR. In: Proceedings of Interspeech, pp. 2645–2648 (2005)
24. Rifkin, R., Schutte, K., Saad, M., Bouvrie, J., Glass, J.: Noise robust phonetic classification with linear regularized least squares and second-order features. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (2007)
25. Raj, B., Turicchia, L., S.-N. B., Sarpeshkar, R.: An FFT-based companding front end for noise-robust automatic speech recognition. In: European Association for Signal Processing Journal on Audio, Speech, and Music Processing, volume 2007 (2007)
26. Hirsch, H.G., Pierce, D.: The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions. Challenges for the Next Millennium, Automatic Speech Recognition (2000)
27. ETSI. ETSI ES 202 050 V1.1.5—Speech Processing, Transmission and Quality Aspects (STQ), Distributed speech recognition, Advanced front-end feature extraction algorithm, Compression algorithms (2007)
28. Lathoud, G., Doss, M., Boulard, H.: Channel normalization for unsupervised spectral subtraction. In: Proceedings of Automatic Speech Recognition and Understanding (2005)
29. Vaseghi, S., Milner, B.: Noise compensation methods for Hidden Markov model speech recognition in adverse environments. *IEEE Trans. Speech Audio Process.* **5**, 11–21 (1997)
30. Martin, R., Breithaupt, C.: Speech enhancement in the DFT domain using Laplacian speech priors. In: Proceedings of International Workshop on Acoustic Echo and Noise, Control (2003)
31. Ephraim, Y., Malah, D.: Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Speech Audio Process.* **32**, 1109–1121 (1984)
32. Grinstead, C., Snell, J.: Introduction to probability. American Mathematical Society, Rhode Island (1997)
33. Dempster, A., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc.* **39**, 1–38 (1977)
34. Moreno, P., Raj, B., Stern, R.: A vector Taylor series approach for environment-independent speech recognition. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 733–736 (1996)
35. Kim, H., Rose, R.: Cepstrum-domain acoustic feature compensation based on decomposition of speech and noise for ASR in noisy environments. *IEEE Trans. Speech Audio Process.* **11**, 435–446 (2003)
36. Deng, J., Bouchard, M., Yeap, T.H.: Noisy speech feature estimation on the Aurora2 database using a switching linear dynamic model. *J. Multimedia* **2**, 47–52 (2007)
37. Windmann, S., Haeb-Umbach, R.: Modeling the dynamics of speech and noise for speech feature enhancement in ASR. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing, pp. 4409–4412 (2008)
38. Li, Y., Erdogan, H., Gao, Y., Marcheret, E.: Incremental on-line feature space MLLR adaptation for telephony speech recognition. In: Proceedings of International Conference on Spoken Language Processing, pp. 1417–1420 (2002)
39. Jankowski, C., Vo, H.-D., Lippmann, R.: A comparison of signal processing front ends for automatic word recognition. *IEEE Trans. Speech Audio Process.* **3**, 286–293 (1995)
40. Kim, J., Kim, L., Hwang, S.: An advanced contrast enhancement using partially overlapped sub-block histogram equalization. *IEEE Trans. Circuits Syst. Video Technol.* **11**, 475–484 (2001)
41. Hilger, F., Ney, H.: Quantile based histogram equalization for noise robust large vocabulary speech recognition. *IEEE Trans. Audio Speech Lang. Process.* **14**, 845–854 (2006)
42. Droppo, J., Deng, L., Acero, A.: A comparison of three non-linear observation models for noisy speech features. In: Proceedings of Eurospeech, vol. 2003, pp. 681–684 (2003)
43. Bar-Shalom, Y., Li, X.: Estimation and Tracking: Principles, Techniques, and Software. Artech House, Norwood (1993)

44. Ganapathiraju, A., Hamaker, J., Picone, J.: Applications of support vector machines to speech recognition. *IEEE Trans. Signal Process.* **52**, 2348–2355 (2004)
45. Bilmes, J.A.: Maximum mutual information based reduction strategies for cross-correlation based joint distributional modeling. In: *Proceedings of ICASSP*, pp. 469–472. Seattle, Washington (1998)
46. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: *Proceedings of International Conference on Machine Learning*, pp. 282–289 (2001)
47. Sha, F., Pereira, F.: Shallow parsing with conditional random fields. In: *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Association for Computational Linguistics, Morristown, NJ, USA, pp. 134–141 (2003)
48. Pinto, D., McCallum, A., Wei, X., Croft, W.: Table extraction using conditional random fields. In: *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 235–242 (2003)
49. Quattoni, A., Collins, M., Darrell, T.: Conditional random fields for object recognition. In: *Advances in Neural Information Processing Systems*, vol. 17, pp. 1097–1104 (2005)
50. Roark, B., Saraclar, M., Collins, M., Johnson, M.: Discriminative language modeling with conditional random fields and the perceptron algorithm. In: *Proceedings of Association for Computational Linguistics*, pp. 48–55 (2004)
51. Schuller, B., Eyben, F., Rigoll, G.: Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech. In: André, E., Dybkjaer, L., Neumann, H., Pieraccini, R., Weber, M. (eds.) *Perception in Multimodal Dialogue Systems: 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*. PIT 2008, Kloster Irsee, Germany, 16–18 June 2008, *Proceedings of Lecture Notes on Computer Science (LNCS)*, vol. 5078, pp. 99–110. Springer, Berlin (2008)
52. Reiter, S., Schuller, B., Rigoll, G.: Hidden conditional random fields for meeting segmentation. In: *Proceedings 8th IEEE International Conference on Multimedia and Expo, ICME 2007*, pp. 639–642, Beijing, China (2007)
53. Rauch, H., Tung, G., Striebel, C.: Maximum likelihood estimates of linear dynamic systems. In: *Journal of American Institute of Aeronautics and Astronautics* vol. 3, pp. 1445–1450 (1965)
54. Barber, D.: Expectation correction for smoothed inference in switching linear dynamical systems. *J. Mach. Learn. Res.* **7**, 2515–2540 (2006)
55. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
56. Fernandez, S., Graves, A., Schmidhuber, J.: An application of recurrent neural networks to discriminative keyword spotting. In: *Proceedings of Internet Corporation for Assigned Names and Numbers 2007*. vol. 4669, pp. 220–229. Porto, Portugal (2007)