# 6

# Bandwidth Extension for Speech

Peter Jax

*Institut für Nachrichtengeräte und Datenverarbeitung (IND);*
*Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen*

In this chapter, the problem of speech enhancement by artificial bandwidth extension is addressed. Whereas in the preceding chapters the signal processing was mostly based on properties of the human auditory system, that is, of the signal sink, the bandwidth extension of speech signals uses properties of the signal source. Hence, here we restrict our view to those bandwidth extension approaches that perform adaptive signal processing according to the well-known time-varying source-filter model of speech production. Note that any of the methods described in the other Chapters 2, 3 and 5 in general can as well be applied to speech signals yet often with lower quality than specialized algorithms due to the lower amount of a priori information that is utilized.

The typical application of bandwidth extension for speech is due to the basic design of speech transmission systems: in current digital public telephone systems the acoustic bandwidth of the transmitted speech signal is usually still limited to the frequency range of the old analogue telephone system, that is, to about 300 Hz to 3.4 kHz. This bandwidth limitation causes the characteristic sound of *telephone speech*.

The minimum requirements on the bandwidth of analogue speech communication systems was specified in the CCITT Red Book from 1961 (see, e.g. Schmidt and Brosze [237]): at the cut-off frequencies of 300 Hz and 3.4 kHz the transmission level may be attenuated by no more than 10 dB with regard to the level at the reference frequency of 800 Hz (ITU-T Rec. G.132 [121], Rec. G.151 [122]). The reasons for the bandwidth limitation at that time were the use of analogue frequency-division multiplex transmission with a frequency grid of 4 kHz, and the optional use of sub-audio telegraphy for out-of-band signalling. The minimum bandwidth of 300 Hz to 3.4 kHz was specified to guarantee an intelligibility of sentences of about 99% from clean telephone speech.

Nowadays, the public telephone system has almost completely been converted to digital transmission techniques. According to the international standard ITU-T Rec. G.711 [123], the speech signals are sampled at a sampling frequency of 8 kHz, and the samples are quantized using the A-law respectively μ-law PCM-encoding laws, yielding a bit rate of 64 kb/s. A strict upper limit of 4 kHz on the transmitted frequency range is enforced by

*Audio Bandwidth Extension: Application of Psychoacoustics, Signal Processing and Loudspeaker Design.* E. Larsen and R. M. Aarts
 ISBN 0-470-85864-8

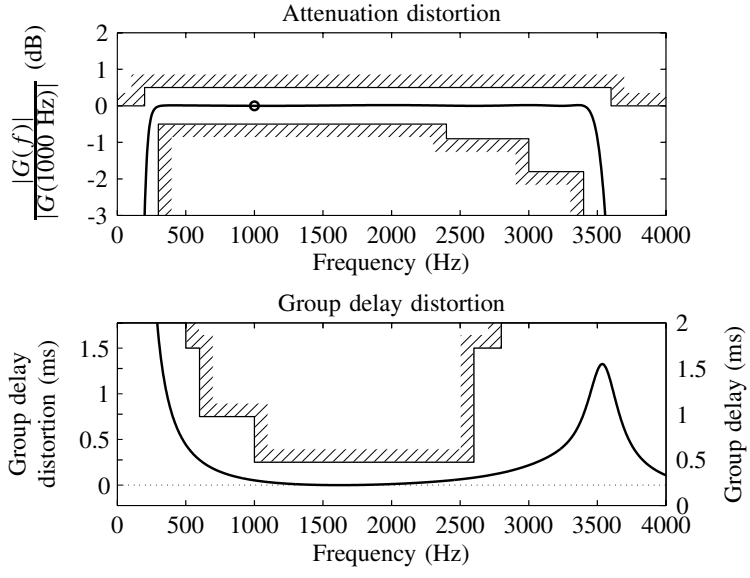


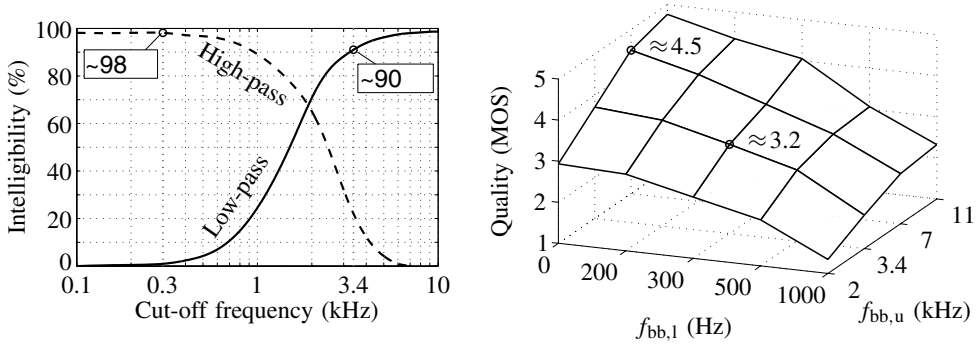

**Figure 6.1** Design constraints from ITU-T Rec. G.712 [124, Sec. 1, 2] for PCM speech transmission. The reference value (0 dB at 1000 Hz) for the attenuation distortion is marked by the filled circle in the upper diagram. The minimum value of the group delay (here: 224 µs at 1621 Hz) is taken as the reference for the group delay distortion. The solid curve gives an example of admissible filter characteristics

the sampling frequency of 8 kHz. Because the implementation of digital circuits in existing networks was performed by successive replacements of analogue circuits, the constraints of the old analogue system applied. The required performance characteristics of PCM transmission channels are specified in detail in the standard ITU-T Rec. G.712 [124]. The design constraints with respect to attenuation and group delay are illustrated in Fig. 6.1.

For mobile radio telephony systems, a further limitation of the frequency range is specified to reduce the amount of disturbing low-pass background noise. In GSM, for example, both the sending and receiving sensitivity of headset or handset mobile terminals shall provide an attenuation of at least 12 dB for low frequencies below 100 Hz (ETSI Rec. GSM 03.50 [68]).

Compared to natural speech, telephone speech has a significantly degraded quality: the removal of low frequencies below about 300 Hz leads to a reduction of the loudness of the speech, leading to a 'thin' voice. In spite of this absence of the fundamental harmonic in the bandlimited speech, a human listener can still perceive the virtual pitch from the harmonic structure of the remaining overtones (Zwicker and Fastl [309], Terhardt [267]), see Sec. 1.4. The elimination of high-frequency components beyond 3.4 kHz, on the other hand, leads to a reduction of the transparency and articulateness of the speech. The bandlimited telephone speech sounds somewhat 'muffled'.

Because both the high- and low-frequency speech components contain some speaker-dependent characteristics, their absence in the bandlimited speech makes it sometimes difficult for a human listener to identify the conversational partner.




(a) Intellig. of meaningless syllables (b) Subjective speech quality

**Figure 6.2** Impacts of a bandwidth limitation on speech intelligibility and subjective quality. In part (a), the intelligibility of meaningless syllables in low-pass respectively high-pass filtered speech is illustrated (data from Terhardt [267]). (b) This compares the speech quality, measured in terms of the subjective *mean opinion score* (MOS), of band-pass-filtered speech with different lower ($f_{bb,l}$) and upper ($f_{bb,u}$) cut-off frequencies (data from Krebber [148])

### Speech Intelligibility

The relevance of high- and low-frequency speech components for the speech intelligibility is pointed out in Fig. 6.2 (a). The diagram shows the intelligibility of (individually) low-pass or high-pass filtered meaningless syllables (French and Steinberg [76], Terhardt [267]). It can be observed that the intelligibility is quite high for the band limits of the telephone band-pass: low-pass filtering with a cut-off frequency of 3.4 kHz yields intelligibilities around 91%, while high-pass filtering at 300 Hz leads to an intelligibility of about 98%.

The intelligibility of meaningless syllables from telephone speech is about 90%, thus making it sometimes necessary to use the spelling alphabet to communicate words that cannot be understood from the context, for example unknown names. The intelligibility of whole *sentences* from clean telephone speech, however, is around 99% (Brosze *et al.* [41], Schmidt and Brosze [237]). Thus, potential benefits of bandwidth extension in terms of the intelligibility of sentences seem to be quite small. Nevertheless, an improvement of the intelligibility of syllables would make the communication more comfortable and less strenuous in many cases, that is, the *listening effort* would be reduced.

### Subjective Speech Quality

Listening experiments have shown that the acoustic bandwidth of speech signals contributes significantly to the perceived speech quality (Krebber [148], Voran [291]). This fact is illustrated in the right diagram of Fig. 6.2 (b), which shows the results of evaluations of subjective speech qualities for clean band-pass-filtered speech. The speech quality is expressed in terms of the *mean opinion score* (MOS), which reflects the subjective rating by human listeners on a scale between one (unacceptable quality) and five

(excellent quality). The two points in Fig. 6.2 (b) that are marked by circles indicate the scores for telephone speech (3.2 MOS points) and 'wideband' speech (4.5 MOS points), respectively.

Starting from the bandwidth of telephone speech (300 Hz to 3.4 kHz), the expansion of the bandwidth both towards low and high frequencies leads to significant gains of the achieved MOS scores. The best scores are obtained by a symmetric expansion towards low *and* high frequencies. In comparison to telephone speech, typical *wideband speech* with a bandwidth of 50 Hz to 7 kHz yields a considerable maximum gain of about 1.3 MOS points.

## 6.1 APPLICATIONS

Owing to the importance of the acoustic bandwidth for speech intelligibility, and especially for the subjective quality, it seems to be worthwhile to aim at an expansion of the transmitted acoustic speech bandwidth. Particularly, in digital communications and hands-free telephony, there is a demand for enhancing the subjective speech quality. True wideband speech communication requires a modification of the transmission link – enhanced speech codecs have to be employed on both sides of the link. Accordingly, several wideband speech-coding schemes have been investigated in the past, aiming at the increase of the acoustic bandwidth to 50 Hz to 7 kHz. In the 1980s, the G.722 codec was standardized by ITU, with bit rates of 64, 56, and 48 kb/s mainly targeting the applications of teleconferencing and ISDN telephony (ITU-T Rec. G.722 [125], Maitre [165]). Later the G.722.1 codec [126] was added with bit rates of 32 and 24 kb/s. Recently, the *adaptive multi-rate wideband* (AMR-WB) codec algorithm (several modes with bit rates from 23.85 down to 6.6 kb/s) was developed and standardized by 3GPP and ETSI [2], Bessette *et al.* [31]. This codec family has also been adopted by the ITU [127]. The implementation of the AMR-WB codec is projected for GSM and 3GPP WCDMA networks.

However, for economical reasons, the bandwidth limitation is not likely to change *on a broad scale* in the near future. It is very likely that, at least for some transitional period, the telephony network will be a mixed network, comprising both narrowband- and wideband-capable terminals.

An alternative approach towards an enhanced acoustic bandwidth of the received speech signal is artificial *bandwidth extension* (BWE) of speech. The challenge of BWE in speech transmission is illustrated in Fig. 6.3: the wideband microphone signal $s_{\mathrm{wb}}$ is band-pass filtered prior to analogue-to-digital conversion and transmitted across the telephone network. At the receiving terminal, only the narrowband signal $s_{\mathrm{nb}}$ is available. This bandlimited speech signal is analysed by the bandwidth extension system. The missing low- and/or high-frequency signal components are estimated and added to the received base-band components. By this, the algorithm determines an estimate $\tilde{s}_{\mathrm{wb}}$ of the wideband speech that is passed on to the loudspeaker.

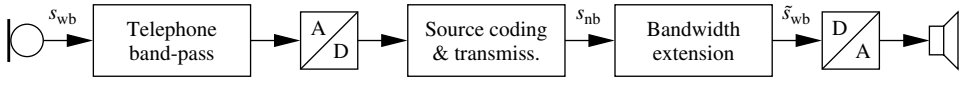



**Figure 6.3** Artificial bandwidth extension in digital speech transmission

The application of bandwidth extension is, in principle, independent of the sending side of the transmission link and of source coding and transmission methods. Hence, the bandwidth extension approach is fully compatible with the existing speech communication infrastructure. It must be emphasized that the concept of artificial bandwidth extension should not be considered to be antagonistic to true wideband coding – on the contrary, it constitutes a harmonious extension to wideband speech services, because it can help reduce the quality variations between the different speech signals in a mixed mode network. Possible fields of application for artificial bandwidth extension systems include the following ones:

- Artificial bandwidth extension can be implemented in a (receiving) terminal equipment as depicted in Fig. 6.3. Then, the user of the terminal gets an improved speech quality, albeit the sending terminal is only capable of narrowband speech transmission. The implementation of bandwidth extension is attractive for manufacturers with respect to the competition on the terminal market.
  It must be noted that there are certain physical constraints caused by the rather small size of modern mobile handsets, particularly for playing low-frequency signals via small loudspeakers (compare Chap. 2). Some loudspeakers have lower cut-off frequencies of up to 1000 Hz, particularly if the small loudspeaker of a mobile phone is operated in hands-free mode. With handsets, the transfer function from the loudspeaker to the ear strongly depends on the positioning of the handset at the ear. If the auricle is not tightly sealed, an acoustic leakage occurs, which impairs the transfer function particularly at low frequencies (Krebber [148]). In many cases, with the aforementioned physical constraints, physical speech bandwidth extension towards *low* frequencies does not make much sense since the extended signal components cannot be provided to the listener.
  For the design of the bandwidth extension algorithm, it should be regarded that, in general, source coding has been applied within transmission. For example, in ISDN the A-/$\mu$-law, PCM-encoding rules from ITU-T Rec. G.711 [123] are used, or in GSM one of the speech codecs specified in ETSI is utilized. It can be observed, however, that coding distortions do not have a major detrimental effect on bandwidth extension, but on the other hand the extension algorithm can benefit from adopting dequantized parameters from the speech decoder.
- In a mixed mode speech communication network, comprising both narrowband- and wideband-capable terminals, artificial bandwidth extension can be implemented within network nodes for transcoding from narrowband codecs to wideband codecs. This is especially beneficial if switchings between narrowband and wideband transmission modes occur, for example, due to handovers in mobile radio access networks [1, Sec. 27].
- If so-called wideband speech (typical frequency range: 50 Hz to 7 kHz) is already available, it is possible to perform bandwidth extension towards 'super-wideband' speech, that is, with a target frequency range of up to 16 kHz. For example, in low bitrate MPEG coding, a special *speech mode* without need to send extra side information as in spectral band replication (SBR) audio coding is possible, or the speech quality of wideband speech codecs can be improved further. This application is even more promising than the extension of telephone speech because the uncertainty of 'super

high frequency' speech components (e.g. between 7 and 16 kHz) is lower and more information can be gained from the available wideband speech signal.

- Bandwidth extension techniques are commonly used within wideband speech codecs, for example, in Dietrich[60], Taori *et al.* [265], in the split-band CELP (SB-CELP) family of speech codecs, such as Paulus [206], Schnitzler [238], Erdmann *et al.* [67]), and in the AMR-WB codec [3]. However, in these approaches mostly the bandwidth extension is applied to a quite narrow frequency band at very high frequencies, for example, only to the signal components between 6 and 7 kHz. Furthermore, the extension can be supported by transmitting side information (compare the *spectral band replication* (SBR) techniques for audio coding in Sec. 5.5).
- One of the first investigated applications of artificial bandwidth extension aimed at the improvement of the quality of telephone contributions in broadcast programmes Croll [54]. If telephone speech is interposed between passages of studio speech, it can become distracting for the listener, because understanding the two different types of speech requires different levels of concentration. By bandwidth extension, the quality of the enhanced speech comes closer to that of studio speech. If the telephone contribution is from a professional correspondent, pre-collected a priori knowledge about the characteristics of the original voice can be made available to the extension algorithm.
- Artificial bandwidth extension can be applied to enhance the acoustical quality of historical recordings of speech. In this application, no real-time processing is required, and the parameters of the algorithm can be tuned manually. If additional wideband recordings of the speaker are available, they can be used to determine the particular voice characteristics.

## 6.2 FROM A SPEECH PRODUCTION MODEL TO THE BANDWIDTH EXTENSION ALGORITHM

In principle, the physical reconstruction of the acoustic bandwidth of (speech) signals can only be feasible if the algorithm has some a priori knowledge about the input signal. For example, if we consider an arbitrary signal that is sampled with a sampling rate of 8 kHz, and if there is no further information available on the kind of the signal, it is impossible due to Nyquist's theorem to tell anything about the signal components beyond the limit frequency of 4 kHz. If, however, a mathematical model of the source of the signal is available, the situation is fundamentally different: both the wideband signal as well as the bandlimited signal are determined by parameters of the common source model. Consequently, exact knowledge of these source parameters would open up the possibility to reconstruct the complete wideband signal as it was originally produced. The parameters of the source, on the other hand, can be estimated from the characteristics of the bandlimited signal.

Because each mathematical model can only be a statistical approximation of the real physical source of a signal, there are several potential drawbacks of such model-based approaches: owing to *simplifications* introduced by the modelling, there will be estimation errors both of the parameters of the source model as well as of the reconstructed wideband signal. In addition, if the characteristics of the actual physical source do not match the

characteristics of the source model exactly, that is, if there is a *model mismatch*, the probability of estimation errors and artefacts in the enhanced signal further increases.

Another cause of estimation errors follows from the basic properties of the signal source. In general, it is not possible to estimate the parameters of the source with arbitrary accuracy because of the random attributes of the signal. In this regard, an upper bound on the quality of the estimation of the spectral envelope of the speech signal will be evaluated in Sec. 6.4.3.

The utilization of a particular model for the signal source also imposes fundamental limits on the application areas of the bandwidth extension algorithm. If, for example, the algorithm is based on a model of the process of speech production, the algorithm will naturally not have the capability to extend general audio signals (such as music), or to reconstruct characteristics of the acoustical environment of the speech signal, such as reverberation or background noise.

### *6.2.1 MODEL OF THE PROCESS OF SPEECH PRODUCTION*

For the process of speech production, there exists a well-known source-filter model, which is illustrated in Fig. 6.4. This model has found wide acceptance in many applications of speech signal processing, especially in the areas of speech coding and speech synthesis (see, e.g. Flanagan [71], Rabiner and Schafer [217], Vary *et al.* [286]). According to the physiology of the human vocal tract apparatus, the model can be sub-divided into two parts: first, an *excitation signal* $u(k)$ is produced, which resembles the excitation of the vocal tract as produced by the vocal cords for voiced sounds, or by a constriction of the

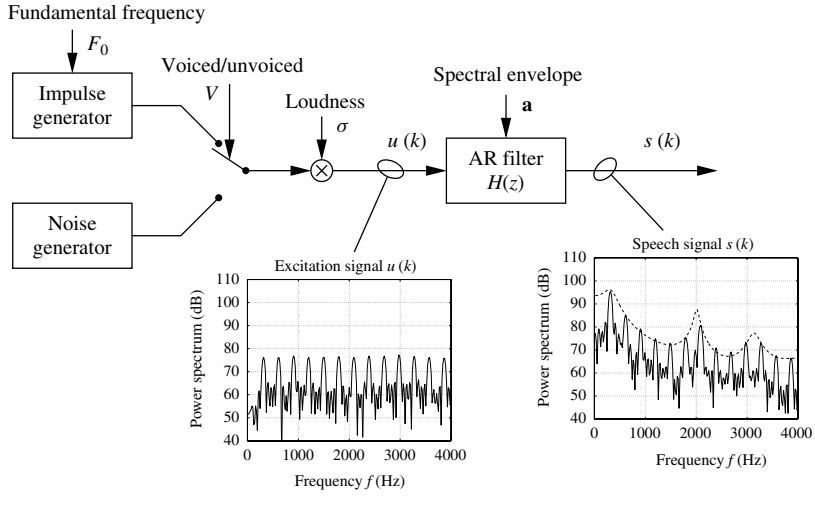



**Figure 6.4** Time-discrete, linear source-filter model of the process of speech production. To clarify the principle, exemplarily the power spectra of the excitation and output signal of the model are shown for an idealized voiced speech sound. The spectral envelope of the speech signal is shaped by the auto-regressive (AR) vocal tract filter $H(z)$. The magnitude transfer function of $H(z)$ is illustrated by the dashed line in the right-hand diagram

vocal tract during unvoiced sounds or plosives. The excitation signal $u(k)$ is the input signal to a purely recursive, that is, auto-regressive (AR), digital filter $H(z)$ that models the resonance characteristics of the vocal tract.

All parameters, $F_0$, $V$, $\sigma$, and $\mathbf{a}$, of the model are basically highly time-variant. However, for speech signals, we can assume that the system is short-term stationary during intervals with a duration of at least 10 to 30 ms. Therefore, in many speech-processing algorithms, the speech signal is processed frame by frame with frame durations between 5 and 30 ms.

***Excitation signal*** In the source-filter model, the excitation signal $u(k)$ is produced by several interacting sub-systems. For voiced sounds (e.g. vowels), a sequence of equidistant impulses with the desired fundamental frequency $F_0$ of the speech signal is produced by an impulse generator. In the frequency domain, this impulse sequence corresponds to several harmonics, which are positioned at the fundamental frequency $F_0$ and integer multiples thereof. All harmonics have the same constant amplitude. An example of the short-term power spectrum of the excitation signal of a voiced sound is shown in the left diagram in Fig. 6.4. For unvoiced sounds, the excitation signal is produced by a noise generator that produces white noise with a variance of 1. Note that for both kinds of excitation, the spectrum of the modelled excitation signal is flat. The selection of the particular kind of excitation is performed by a binary switch that is controlled by the voicing parameter $V$. Finally, the gain of the speech signal is specified with the common scalar gain factor $\sigma$.

This simple model of the generation of the excitation signal reflects the real physical process speech production in a very idealized and simplified manner. For example, it is very rare that speech sounds are exclusively of voiced or unvoiced nature. Normally, the excitation signal is a mixture of both kinds of excitation. Further, the excitation of the human vocal tract is not perfectly spectrally flat in reality: the periodic excitation produced by the vocal cords in general has low-pass characteristics; for unvoiced sounds, on the other hand, the spectral characteristics of the excitation signal depend on shape and position of the constriction in the vocal tract that causes the chaotic turbulences of the air. However, such model mismatches of the spectral characteristics of the excitation signal can be taken into account by the subsequent filter $H(z)$ in the source-filter model. Although the model, in a sense, is too simple to describe the complex physical mechanism of speech production, it has proven to be sufficient for most applications of speech processing.

***Vocal tract filter*** In the human vocal tract, the sound-specific spectral envelope of the speech signal is shaped. Its signal-processing model consists of a time-variant auto-regressive (AR) filter

$$H(z) = \frac{1}{A(z)}, \quad \text{with } A(z) = \sum_{i=0}^{N_a} a_i \, z^{-i}. \tag{6.1}$$

The purely recursive structure of the filter $H(z)$ can be motivated by physically modelling the human vocal tract via an idealized, that is, lossless and discretized, acoustic tube with varying diameter (e.g. Flanagan [71], Rabiner and Schafer [217], Vary *et al.* [286]).

Because the human ear is basically insensitive to moderate variations of the signal phase, the vocal tract can be modelled by a minimum-phase AR filter.

According to its role in the decoder of a *linear predictive coding* (LPC) system, the filter modelling the vocal tract is frequently called *LP synthesis filter* in literature. The filter coefficients are combined in the column vector $\mathbf{a} = \left[a_0, a_1, \ldots, a_{N_a}\right]^{\mathrm{T}}$. The first coefficient is normalized to $a_0 = 1$ in general such that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{A(\mathrm{e}^{j\Omega})}\,\mathrm{d}\Omega = \frac{1}{2\pi}\int_{-\pi}^{\pi}A(\mathrm{e}^{j\Omega})\,\mathrm{d}\Omega = 1. \tag{6.2}$$

Owing to this normalization, the transfer function of the vocal tract filter is independent of the short-term power (or gain) of the speech signal. Hence, it describes the *shape* of the spectral envelope only. Because of the limited order $N_a$ of the AR filter, it describes a smoothed version of the spectral envelope of the signal. Typical filter orders are $N_a = 8\ldots10$ for narrowband speech (sampling frequency $f_\mathrm{s} = 8\,\mathrm{kHz}$), and $N_a = 16\ldots18$ for wideband speech ($f_\mathrm{s} = 16$ kHz).

Since the LP synthesis filter $H(z)$ is a minimum-phase filter, there always exists a stable inverse thereof. The inverse $A(z)$ of the LP synthesis filter has a finite impulse response, and it is often called *LP analysis filter*. The analysis filter has an important property for the bandwidth extension application: if the filter coefficients $\mathbf{a}$ are available, that is, if the shape of the spectral envelope of the speech signal is known, applying the LP analysis filter to the speech signal will calculate an estimate of the excitation signal $u(k)$.

The optimal filter coefficients (in the sense of minimizing the power of the estimated excitation signal) for a given segment of speech can be determined by performing a linear prediction analysis of the speech frame (see, e.g. Makhoul [166], Markel and Gray [168], Vary *et al.* [286]). This procedure is commonly performed in two stages: for each speech segment the first $N_a + 1$ short-term auto-correlation coefficients are estimated, which are then transformed into the filter coefficients $\mathbf{a}$, for example, by the recursive Levinson–Durbin algorithm (Roberts and Mullis [228]).

### *6.2.2 BANDWIDTH EXTENSION ALGORITHM*

Most adaptive bandwidth extension algorithms for speech are based on the source-filter model of the speech production process as described in the previous section. The estimation of the missing signal components is performed in a two-stage procedure, indirectly via the model of the source: in the first step the parameters of the wideband source model are estimated from the bandlimited speech signal. These parameters are then used in combination with the model itself to determine an estimate of the wideband speech. This approach is in general well suited for the extension both to high frequencies and to low frequencies.

Below, on the basis of the block diagram in Fig. 6.5, an overview of the principal structure and properties of bandwidth extension algorithms for speech shall be given. According to the structure of the source-filter model from Sec. 6.2.1, the bandwidth extension is performed separately for the excitation signal and for the spectral envelope of the speech signal (Cheng *et al.* [49], Carl [44], Iyengar *et al.*, see Chapter 8). Since these two constituents of the speech signal can be assumed to be mutually independent

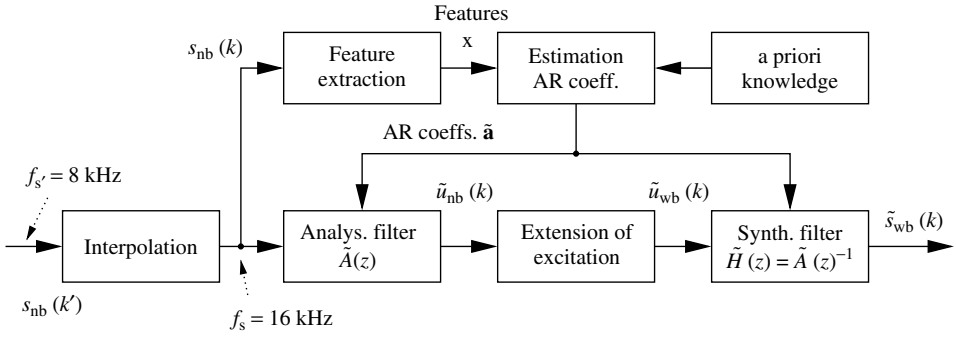



**Figure 6.5** Signal flow of an algorithm for the bandwidth extension of speech signals (Jax *et al.* [129]). The final synthesis filter $\tilde{H}(z)$ of the algorithm and certain parts of the sub-system for the extension of the excitation signal reflect the source-filter model from Sec. 6.2.1

to a certain extent, the separate optimization of the two parts of the algorithm leads to an approximation of the global optimum.

The importances of the two sub-tasks are different. For the extension towards high frequencies, the principal problem is posed by the estimation of the wideband spectral envelope. This fact can be verified easily in listening experiments by applying the BWE algorithm utilizing knowledge of the original wideband spectral envelope: by modifying only the excitation signal in the extended frequency bands, the quality of the enhanced speech signal is only slightly inferior to the quality of the original wideband speech (see Sec. 6.3 or Carl [44]). Consequently, the sub-system for the estimation of the spectral envelope has to be designed with special diligence.

For low-frequency BWE, an additional important problem is the correct reconstruction of the pitch information. If the fundamental frequency and/or first overtones thereof are recovered incorrectly, the base-band and extended components will not be grouped to a single auditory stream, see Sec. 1.4.

A detailed description of the two parts of the algorithm concerned with the extension of the excitation signal and of the spectral envelope can be found in Secs. 6.3 and 6.4 ff. respectively.

***Interpolation*** If the sampling rate of the input signal $s_{nb}(k')$ of the BWE algorithm is not sufficiently high to allow the representation of the extended speech signal, the first step in the BWE system consists of increasing the sampling rate via interpolation (e.g. Oetken and Schüßler [189], Crochiere and Rabiner [53]). In the example that is illustrated in Fig. 6.5, the narrowband input signal is represented with the typical sampling rate (of narrowband speech) of $f_{s'} = 8\,\text{kHz}$. Hence, to allow the extension of high-frequency components up to a cut-off frequency of 7 kHz, the sampling rate has to be increased to $f_s = 16\,\text{kHz}$. Note that by the interpolation the signal contents are not modified – the interpolated signal $s_{nb}(k)$ is still bandlimited in the same manner as the input signal $s_{nb}(k')$.

All of the subsequent modules are processed with the fixed sampling rate $f_s$, for example, $f_s = 16\,\text{kHz}$. Furthermore, the processing is mostly performed frame by frame with a frame length of about 20 ms. In the sequel, the frame index is denoted by $m$. Within each frame, the samples are indexed by the variable $\kappa$, with $0 \leq \kappa \leq N_\kappa - 1$, and $N_\kappa$ being the number of samples per frame (i.e. $N_\kappa = 320$ if $f_s = 16$ kHz).

***Estimation of the AR coefficients*** The actual bandwidth extension starts with the estimation of the coefficient set $\tilde{\mathbf{a}}$, representing the shape of the spectral envelope of the *wideband* speech signal. For this purpose, as much relevant information as possible shall be utilized from the available bandlimited speech. For each signal frame, a vector of features $\mathbf{x}$ of the input speech is calculated, providing the basis for the estimation. A pre-trained statistical model contributes the necessary a priori knowledge on the properties of the process of speech production. A detailed description of the statistical modelling and of different estimation procedures is given in the Secs. 6.4 to 6.9.

***Analysis filter*** The estimated wideband filter coefficient set $\tilde{\mathbf{a}}$ is utilized in an FIR analysis filter $\tilde{A}(z)$, which is applied to the interpolated bandlimited input signal $s_{\text{nb}}(k)$:

$$\tilde{A}(z) = \sum_{i=0}^{N_a} \tilde{a}_i \, z^{-i}, \quad \text{and} \quad \tilde{u}_{\text{nb}}(k) = \sum_{i=0}^{N_a} \tilde{a}_i \, s_{\text{nb}}(k-i). \tag{6.3}$$

Because the analysis filter is the inverse of the corresponding auto-regressive vocal tract filter, the output $\tilde{u}_{\text{nb}}(k)$ of the analysis filter can be interpreted as an approximation of the excitation of the speech. It must be kept in mind, however, that this estimate is bandlimited in the same manner as the input signal of the BWE algorithm.

***Extension of the excitation signal*** The next step in the BWE system consists of substituting the missing frequency components in the excitation signal. Depending on the desired quality of the extended excitation signal, as well as on the admissible complexity of this sub-system, the different parameters, $\sigma$, $V$, or $F_0$, of the source model can be considered to a greater or lesser extent for this purpose. Owing to the assumed spectral flatness of the excitation signal, and because of the fact that the human ear is quite insensitive to variations of the spectral fine structure at high frequencies, the extension can be realized in a very efficient manner. Different approaches for the extension of the excitation signal are described in Sec. 6.3.

In principle, an extension of low (e.g. below 300 Hz) as well as high components (above 3.4 kHz) of the excitation signal is obtainable. Therefore, the output signal $\tilde{u}_{\text{wb}}(k)$ of this block reflects the desired estimate of the wideband excitation signal.

During the extension of the excitation, it shall be guaranteed that the base-band components of $\tilde{u}_{\text{nb}}(k)$ are not modified – then, the input speech $s_{\text{nb}}(k)$ will be contained transparently in the output signal $\tilde{s}_{\text{wb}}(k)$ of the BWE system.

***Synthesis filter*** So far, both an estimate $\tilde{u}_{\text{wb}}(k)$ of the wideband excitation signal and an approximation $\tilde{\mathbf{a}}$ of the coefficient set of the AR filter representing the spectral envelope

of the wideband speech signal have been determined. To finalize the estimate of the wideband speech signal, the two quantities are combined by means of the synthesis filter

$$\tilde{H}(z) = \left( \sum_{i=0}^{N_a} \tilde{a}_i \, z^{-i} \right)^{-1} = \frac{1}{\tilde{A}(z)}. \tag{6.4}$$

Considering the normalization ($\tilde{a}_0 = 1$) of the AR coefficients, the output signal of the bandwidth extension system is computed by

$$\tilde{s}_{\mathrm{wb}}(k) = \tilde{u}_{\mathrm{wb}}(k) - \sum_{i=1}^{N_a} \tilde{a}_i \, \tilde{s}_{\mathrm{wb}}(k-i). \tag{6.5}$$

Note that the transfer function of the synthesis filter is inverse to the transfer function of the employed analysis filter for each signal frame, because the identical coefficient set $\tilde{\mathbf{a}}$ is utilized in both filters.

### *6.2.3 ALTERNATIVE STRUCTURES*

A characteristic property of the algorithm from Sec. 6.2.2 is the fact that both the analysis filter and the synthesis filter are operated at the same sampling rate and, moreover, with the identical coefficient set $\tilde{\mathbf{a}}$. The two filters are exactly mutually inverse. This feature discriminates the algorithm from alternative approaches, which, in a similar manner, perform a separate extension of the excitation signal and the spectral envelope (e.g. Carl [44, 45], Avenando *et al.* [23], Nakatoh *et al.* [183], Enbom and Kleijn [64], Epps and Holmes [66], Miet *et al.* [174], Park and Kim [200], Valin and Lefebvre [279], Fuemmeler *et al.* [78]). In the latter algorithms, narrowband coefficients $\tilde{\mathbf{a}}'_{\mathrm{nb}}$ for the analysis filter are either determined via an LP analysis or taken from a codebook. For the synthesis filter, on the other hand, a different wideband coefficient set $\tilde{\mathbf{a}}_{\mathrm{wb}}$ is utilized, which is estimated or taken from a different codebook (the so-called shadow-codebookcodebook,shadow-codebook). In Fig. 6.6, an example of the structure of such an alternative BWE algorithm is shown. Note that because with the structure from Fig. 6.6 the LP analysis and synthesis filters are not exactly mutually inverse in the base-band, it is important to apply gain correction of the extended speech components.

The two basic approaches from Figs. 6.5 and 6.6 have distinct properties that shall be discussed in the following text.

***Transparency in the base-band and mixing*** An important requirement on algorithms for the bandwidth extension of speech signals is the transparency of the system with respect to the bandlimited input signal. As the input signal provides the best possible speech quality within its limited frequency range, it shall be contained unmodified in the output of the BWE system.

If different coefficient sets (as, e.g., $\tilde{\mathbf{a}}'_{\mathrm{nb}}$ and $\tilde{\mathbf{a}}_{\mathrm{wb}}$ in Fig. 6.6) are used in the analysis and synthesis filter, it is in general necessary, in order to ensure the transparency of the algorithm with respect to the base-band, to mix the original input signal of the BWE algorithm with the band-stop-filtered extended speech to calculate the output signal of

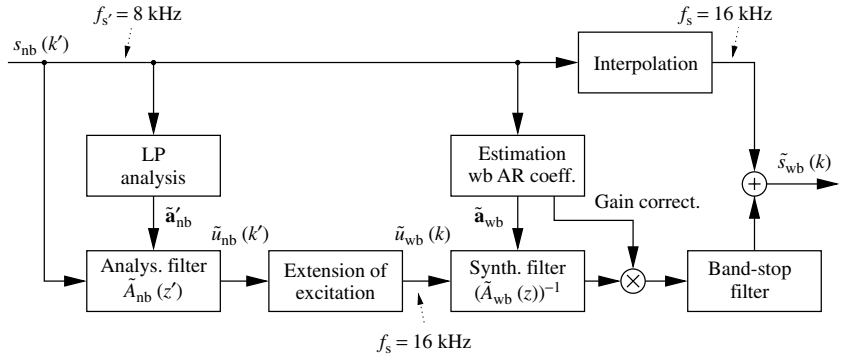



**Figure 6.6** Signal flow of an alternative algorithm for bandwidth extension of speech signals (e.g. Park and Kim [200], Fuemmeler *et al.* [78]). The block 'extension of excitation' additionally performs an interpolation of the excitation signal

the BWE system (see Fig. 6.6). Owing to discrepancies in the transfer functions of the analysis and synthesis filter, the base-band speech components are distorted in the lower signal path of Fig. 6.6, and additionally certain artefacts are produced. Further, the relative power of the extended speech is generally altered with this structure. Therefore, prior to mixing the extended signal with the input speech, a correction factor has to be applied to the extended signal (Carl [44], Park and Kim [200], Nilsson and Kleijn [187], Fuemmeler *et al.* [78]). The proper correction factors have to be estimated in addition to the wideband spectral envelope of the speech.

Such measures are not needed in the algorithm from Sec. 6.2.2 because synthesis and analysis filters are mutually inverse: the transparency of the BWE system gets independent of the extension of the spectral envelope. Provided that during the extension of the excitation the narrowband components of the excitation signal are not modified, errors in the estimated spectral envelope of the analysis filter (in the sense of an optimal LPC prediction filtering) are completely compensated by the inverse synthesis filter *within the base-band*. The power of the signal is not modified because of the filtering. However, to achieve transparency of the complete BWE system, the sub-system that is responsible for the extension of the excitation is now required to be transparent with respect to the base-band components (cf. Sec. 6.3).

***Impact of estimation errors of the spectral envelope*** If the estimated coefficient set $\tilde{\mathbf{a}}$, representing the spectral envelope of the wideband speech, is inaccurate, the impact on the quality of the extended frequency bands of the output signal may be two-fold. It is quite obvious that errors of the frequency response of the synthesis filter within the extended frequency bands directly effect the quality of the extended bands: the estimated excitation signal will be shaped by the synthesis filter according to the erroneous spectral envelope.

In the algorithm from Sec. 6.2.2, errors within the base-band of the frequency response of the analysis filter can further impair the quality of the extended bands in an indirect manner. Owing to errors of $\tilde{\mathbf{a}}$, the estimate $\tilde{u}_{nb}(k)$ of the bandlimited excitation signal, that is determined by the erroneous analysis filter, is not spectrally flat as assumed in

the source-filter model. During the subsequent extension of the excitation signal, these errors within $\tilde{u}_{\mathrm{nb}}(k)$ will propagate into the extended frequency bands of the estimated wideband excitation $\tilde{u}_{\mathrm{wb}}(k)$. Thus, although base-band transparency is guaranteed by the algorithm, errors in the base-band of the estimated spectral envelope do, nevertheless, effect the extended speech signal $\tilde{s}_{\mathrm{wb}}(k)$. If, for example, the extension of the excitation is performed by spectral translation or folding (see Sec. 6.3.3), the errors of the estimated spectral envelope within the base-band and within the extended band are effectively added up. In Sec. 6.4.1.1, a method that prevents errors in the base-band of the estimated spectral envelope will be described.

***Algorithmic delay*** In real-time speech communication, it is generally desired to keep the signal delay as low as possible. Nevertheless, it is important to apply proper delay compensation in any parallel signal path of a BWE system to ensure that the extended frequency components are psychoacoustically grouped together with the base-band speech (cf. Sec. 1.4).

There are several potential sources of algorithmic delay in the bandwidth extension systems of Figs. 6.5 and 6.6. Firstly, there is always an algorithmic delay due to the frame-based processing of the algorithm: all $N_\kappa$ samples have to be available before the processing of a frame can start. If the bandwidth extension algorithm is positioned behind a speech decoder, however, in general this source of delay is not relevant since most speech codecs also operate on a per-frame basis. The bandwidth extension system can be merged with the speech decoder.

If the input signal of the BWE algorithm has to be interpolated before applying the analysis filter, an additional delay will be caused by the interpolation low-pass filter. The design criteria of transition bandwidth and stopband attenuation for this filter, however, are not as stringent as for an isolated interpolation system. The high-frequency part of the speech signal will be approximated anyhow by the BWE system. Aliasing errors from non-optimal interpolation may be masked by subsequently added extended frequency components. Consequently, the order and delay of the interpolation filter can be kept rather low.

A further delay of the speech signal might be introduced if any filters are utilized in the sub-system for the extension of the excitation signal to guarantee the base-band transparency of that sub-system (compare Sec. 6.3). Note that in this case also the adjustment of the synthesis filter coefficient set $\tilde{\mathbf{a}}$ has to be delayed accordingly. The processing of the speech signal by the analysis and synthesis filters does not produce any delay of the signal (although both filters are causal) because the two filters are both minimum-phase filters and mutually inverse.

Finally, a delay of the speech signal is necessary if a look-ahead shall be utilized in the estimation of the wideband spectral envelope (compare Sec. 6.9). In this case, the input signal $s_{\mathrm{nb}}(k)$ of the analysis filter has to be delayed in accordance with the implicit delay of the estimated AR coefficients $\tilde{\mathbf{a}}$.

## 6.3 EXTENSION OF THE EXCITATION SIGNAL

In this section, the sub-system of the BWE algorithm that is responsible for the extension of the excitation signal of the speech (compare Fig. 6.5) is treated. This sub-system gets

the bandlimited estimate $\tilde{u}_{\mathrm{nb}}(k)$ of the excitation as its input. The output signal $\tilde{u}_{\mathrm{wb}}(k)$ on the other hand serves as the input to the final synthesis filter of the BWE system and reflects an estimate of the wideband excitation signal. The task of the extension of the excitation signal is the recovery of the *spectral fine structure* of the speech signal.

Potential algorithms that can be employed for the extension of the excitation signal benefit both from the quite simple structure of the excitation signal according to the source-filter model of speech production (compare Sec. 6.2.1) as well as from insensitivities of the human auditory system with regard to certain distortions of the spectral fine structure at high respectively low frequencies. In this chapter, several algorithms from literature are described and evaluated. The different methods for the extension of the excitation either reuse the signal components of the estimated bandlimited excitation signal $\tilde{u}_{\mathrm{nb}}(k)$, for example, by spectral translation (Sec. 6.3.3) or pitch scaling (Sec. 6.3.4), or they generate new components via explicit signal generation (Sec. 6.3.1) or by non-linear distortion (Sec. 6.3.2).

An important requirement that has to be demanded for the estimated wideband excitation signal $\tilde{u}_{\mathrm{wb}}(k)$ is that it transparently contains the estimated bandlimited excitation signal $\tilde{u}_{\mathrm{nb}}(k)$ – in this case, the complete BWE system becomes transparent with respect to the narrowband input speech (see Sec. 6.2.3). To guarantee this transparency, it is necessary for some of the following methods to mix the original bandlimited excitation $\tilde{u}_{\mathrm{nb}}(k)$ with an appropriately high-pass respectively low-pass-filtered version of the extended excitation.

Because the vast majority of publications on the topic of bandwidth extension of speech signals to date is concerned primarily with the extension of the spectral envelope of the speech, most of the known methods for the extension of the excitation signal have been adopted from the field of speech coding. Especially, techniques from so-called *base-band codecs* are used. In these speech codecs, only a part of the frequency components of the LPC residual signal is coded and transmitted while the remaining components are recovered at the receiving site via *high-frequency regeneration* (HFR, e.g. Makhoul and Berouti [167], Kroon *et al.* [150], Taori *et al.* [265], McCree *et al.* [171]). A prominent representative of this category of speech codecs is the GSM full-rate codec [69], Vary *et al.* [285].

### *6.3.1 EXPLICIT SIGNAL GENERATION*

The most straightforward solution to extend the excitation signal consists of the explicit generation of the missing signal components. Basically, by this approach the excitation part of the source model from Sec. 6.2.1 is implemented directly. The method therefore strongly depends on estimates of the source parameters, that is, on estimates of the voiced/unvoiced state $V$, the gain factor $\sigma$, and the fundamental frequency $F_0$ of the speech (compare Fig. 6.7).

According to the admissible computational complexity and to the desired accuracy of the simulated source model, there are several prevalent approaches:

- *Noise only*: The missing components of the excitation signal are produced by a noise generator and a subsequent band-stop filter. Further, the gain $\tilde{\sigma}$ of the noise signal $\tilde{u}_{\mathrm{mb}}(k)$ has to be adapted to match the gain of the base-band excitation signal $\tilde{u}_{\mathrm{nb}}(k)$. The procedure is reflected by the block diagram of Fig. 6.7 if the voiced/unvoiced

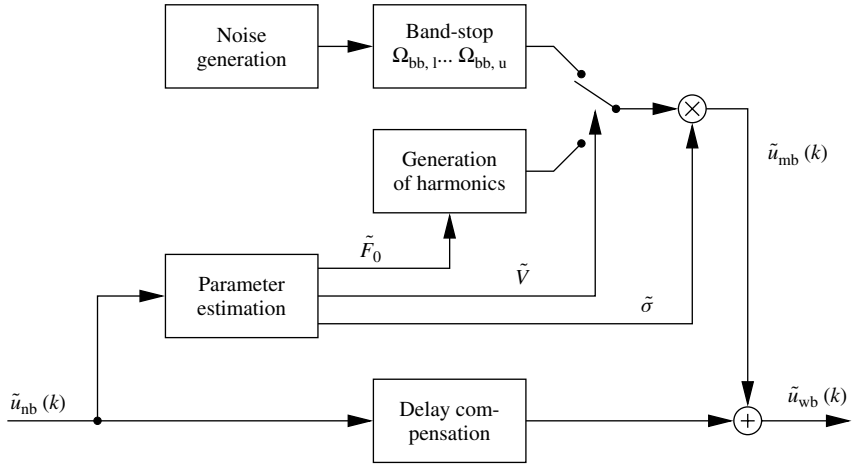



**Figure 6.7** Extension of the excitation signal via explicit signal generation

switch (parameter $\tilde{V}$) is invariably set to the upper (unvoiced) position. This approach can be motivated by the fact that the main contribution of high-frequency components of the speech is during unvoiced sounds.

In fact, the addition of noise components yields very good results if the extended frequency band is rather narrow. For example, this approach has been used successfully in several wideband speech codecs for the coding of the frequency components above 6 or 6.4 kHz and up to 7 kHz, for example, Paulus [206], Schnitzler [238], Erdmann *et al.* [67], 3GPP TS 26.190 [3]. Regarding the quite wide missing frequency band above 3.4 kHz in the artificial bandwidth extension of narrowband telephone speech, however, both the extended excitation signal as well as the extended speech signal (after the AR synthesis filter) sound quite noisy. Especially during voiced speech segments, the noisy signal components added at high frequencies are then well audible and annoying.

- *Noise and/or sinusoids*: The algorithm can be refined by distinguishing between voiced and unvoiced segments of the speech. During unvoiced phases a noise generator is utilized, and during voiced sounds a tonal excitation is produced in the missing frequency band. The techniques for sine-wave generation resemble those from *sinusoidal* or *harmonic coding* (Griffin and Lim [100], Carl [44], McAulay and Quatieri [170]). The voiced/unvoiced switching can either be 'hard', allowing either a noisy or a tonal extended excitation $\tilde{u}_{mb}(k)$ at a time, or the kind of excitation is specified individually for different frequency bands. In the latter case, the approach corresponds to the *harmonic plus noise model* (HNM) from Griffin and Lim [100], Abrantes *et al.* [13], Stylianou [258]. In the bandwidth extension system, the newly generated signal components $\tilde{u}_{mb}(k)$ are mixed with the original bandlimited excitation $\tilde{u}_{nb}(k)$.

  In informal listening experiments, it can be found that a very good estimation of the fundamental speech frequency $F_0$ is crucial for the generation of tonal speech components: if the estimate $\tilde{F}_0$ is inaccurate, the objectionable impression is produced that an interfering simultaneous speaker with a slightly different pitch frequency is added to the speech signal. This problem can be circumvented if the excitation signal

is substituted completely, that is, if also the base-band components are regenerated (Chan and Hui [47, 48]). This, however, may also introduce artefacts in the base-band frequency range of $\tilde{s}_{\mathrm{wb}}(k)$.

### *6.3.2 NON-LINEAR PROCESSING*

The first approach to the artificial bandwidth extension of speech signals to our knowledge was the application of non-linear distortions to the narrowband speech $s_{\mathrm{nb}}(k)$ as proposed by Schmidt [236]. The same basic method can also be used to extend the excitation signal of the speech: an estimate of the wideband excitation signal is determined by applying a non-linear function $g(\cdot)$ to the bandlimited excitation $\tilde{u}_{\mathrm{nb}}(k)$

$$\tilde{u}_{\mathrm{nl}}(k) = g\left(\tilde{u}_{\mathrm{nb}}(k)\right). \tag{6.6}$$

Owing to the non-linear processing, (harmonic) distortions that reflect the desired new signal components in the missing frequency bands are created. The signal $\tilde{u}_{\mathrm{nl}}(k)$ denotes a generalized extended excitation signal here, that is, it can correspond to the high- or low-frequency band of the speech signal, respectively.

There is an unlimited number of possible non-linear functions $g(\cdot)$, and it is quite difficult to find that particular function that yields the best results in the bandwidth extension application. Non-linear functions have been used in bandwidth extension algorithms mainly for the generation of low-frequency speech components to date (Schmidt [236], Croll [54], Patrick *et al.* [201], Valin and Lefebvre [279], Kornagel [147]). The utilized non-linearities $g(\cdot)$ have been, for example, quadratic, cubic, or saturation functions, and half-wave respectively full-wave rectification. Note that here in contrast to Chapters 2 and 3 the non-linearities are applied to a signal containing more than only one harmonic. Unfortunately, the effects of the non-linear function $g(\cdot)$ are very difficult to predict, as any modification of the input signal $\tilde{u}_{\mathrm{nb}}(k)$ (e.g. scaling, the addition of signal components, application of phase distortions, or a simple addition of a constant value) can significantly effect the properties of the distorted signal $\tilde{u}_{\mathrm{nl}}(k)$. In general, either the narrowband signal $\tilde{u}_{\mathrm{nb}}(k)$ has to be pre-processed (normalized) prior to applying the non-linearity, or the distorted signal $\tilde{u}_{\mathrm{nl}}(k)$ has to be post-processed.

The possibility of a post-processing of the distorted signal is illustrated in Fig. 6.8. The shape of the envelope as well as the gain of the distorted signal $\tilde{u}_{\mathrm{nl}}(k)$ are corrected (and adapted to the base-band excitation) to match the assumption of a spectrally flat wideband excitation signal from the source-filter model of Sec. 6.2.1: first, an LP analysis of the signal is performed. The corresponding adaptive LP analysis filter is applied, yielding a whitening of the signal. Afterwards, the gain of the signal is adjusted to match the gain of the base-band excitation $\tilde{u}_{\mathrm{nb}}(k)$. As the non-linear distortion effects the whole frequency range of the distorted signal, measures have to be taken to guarantee base-band transparency. The base-band components have to be removed from $\tilde{u}_{\mathrm{eb}}(k)$ by band-stop filtering, or by high-pass respectively low-pass filtering if only an extension towards high respectively low frequencies is desired. The filtered signal $\tilde{u}_{\mathrm{mb}}(k)$ is mixed with the narrowband excitation signal to determine the estimated wideband excitation signal $\tilde{u}_{\mathrm{wb}}(k)$.

In experiments, very good results were achieved with a simple quadratic non-linear distortion, that is, with the function $\tilde{u}_{\mathrm{nl}}(k) = (\tilde{u}_{\mathrm{nb}}(k))^2$. By this non-linearity, both

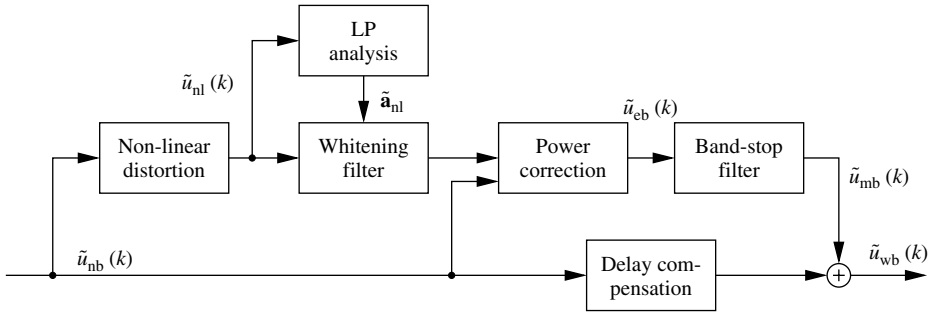



**Figure 6.8** Extension of the excitation signal by non-linear distortion. To compensate for hardly controllable adverse effects of the non-linearity, sophisticated post-processing is required. The algorithmic delay of the band-stop filter has to be compensated for in the path of the base-band signal

low-frequency and high-frequency components can be generated. The upper cut-off frequency of the distorted signal $\tilde{u}_{\mathrm{nl}}(k)$ is twice the upper band limit of the narrowband excitation $\tilde{u}_{\mathrm{nb}}(k)$. Since only *harmonic* distortions are produced by the squaring operation, the tonal components in the enhanced excitation signal $\tilde{u}_{\mathrm{wb}}(k)$ always match the harmonic structure of the bandlimited excitation $\tilde{u}_{\mathrm{nb}}(k)$ during voiced sounds.

For low-frequency bandwidth extension, simple full-wave rectification has proven successful. It has the advantage that the signal level is not altered (a full-wave rectifier is a homogeneous system, see Sec. 1.1.1) such that it can easily be implemented.

### *6.3.3 MODULATION IN THE TIME DOMAIN*

In this section, we consider algorithms that are based on a modulation of the bandlimited excitation signal $\tilde{u}_{\mathrm{nb}}(k)$ (Carl [44], Fuemmeler *et al.* [77], Kornagel [147], Jax and Vary [130]). Because a modulation in the time domain corresponds to a translation in the frequency domain, the input signal is virtually reused by 'shifting' it into the missing frequency band(s). Several well-known methods for the extension of the excitation signal, such as spectral folding or spectral translation, are special cases of the more general modulation concept (Carl [44]).

The straightforward implementation of a spectral translation would be based on the analytical signal of the bandlimited excitation. The product of the analytic signal with a complex-valued modulation function directly yields the desired extended signal. However, the determination of the analytic signal by Hilbert transformation either in the time- or frequency domain is quite complex (e.g. Schuessler [242], Marple [169]). In general, equivalent results as with the analytic signal can also be achieved by modulation of the input signal with a *real-valued* modulation function. However, in this case the shifted spectra cause mutual overlappings that have to be removed by subsequent frequency-selective filtering as illustrated in Fig. 6.9.

In the following, the modulation shall be performed using a real-valued cosine function:

$$\tilde{u}_{\mathrm{M}}(k) = \tilde{u}_{\mathrm{nb}}(k) \cdot \zeta \, \cos(\Omega_{\mathrm{M}} k). \tag{6.7}$$

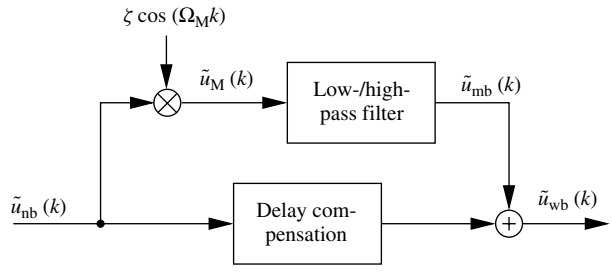



**Figure 6.9** Extension of the excitation signal by modulation. The algorithmic delay of the high-pass filter has to be compensated for in the path of the base-band signal $\tilde{u}_{\mathrm{nb}}(k)$

Depending on the particular modulation frequency $\Omega_{\mathrm{M}}$, the fixed scalar factor $\zeta$ in Eqn. 6.7 has to be chosen from $\zeta \in \{1, 2\}$ to obtain the correct power of the extended excitation signal. The argument of the cosine function consists of the phase function $\Omega_{\mathrm{M}}\, k$, which is linear in time if the modulation frequency $\Omega_{\mathrm{M}}$ is fixed. By the multiplication of the input signal with the cosine signal in time domain, two shifted copies of the original spectrum $\tilde{U}_{\mathrm{nb}}(e^{j\Omega})$ are generated

$$\tilde{U}_{\mathrm{M}}(e^{j\Omega}) = \frac{\zeta}{2}(\tilde{U}_{\mathrm{nb}}(\mathrm{e}^{j(\Omega-\Omega_{\mathrm{M}})}) + \tilde{U}_{\mathrm{nb}}(\mathrm{e}^{j(\Omega+\Omega_{\mathrm{M}})})). \tag{6.8}$$

The two shifted spectra may overlap in different frequency ranges. Whether such overlappings occur, and to which extent, depends on the lower- and upper-frequency limits $\Omega_{\mathrm{bb,l}}$ and $\Omega_{\mathrm{bb,u}}$ of the bandlimited speech signal as well as on the modulation frequency $\Omega_{\mathrm{M}}$.

The modulation approach to the extension of the excitation signal is especially suited for the extension of high frequencies, because the frequency range to be extended above the upper band limit $\Omega_{\mathrm{bb,u}}$ of the input speech is – in contrast to the lower extended frequency range – in general larger than the bandwidth of the base-band. This property has consequences for the design criteria of the high-pass respectively low-pass filter from Fig. 6.9: if the bandwidth of the shifted spectrum is greater than the width of the extended frequency range, the implemented filter shall have very steep slopes.

Owing to the importance of the fundamental frequency of the speech in the low-frequency excitation, only a pitch-adaptive approach is applicable to the extension of the missing low-frequency band.

#### 6.3.3.1 Spectral Folding

The method of spectral folding reflects a special case of the modulation method that is exclusively suitable for the extension of high-frequency components. The modulation frequency is specified to be equal to the Nyquist frequency $\Omega_{\mathrm{M}} = \pi$ (corresponding to $f_{\mathrm{M}} = 8\,\mathrm{kHz}$ for the (interpolated) excitation signal $\tilde{u}_{\mathrm{nb}}(k)$). Thereby, the modulation function is simplified to a sequence of alternating signs $\cos(\Omega_{\mathrm{M}}\, k) = (-1)^k$. The two shifted spectra are superimposed constructively such that $\zeta = 1$. By the modulation, a 'folded' version of the input signal is generated, the spectrum of which is mirrored at $\Omega = \pi/2$, that is, the half of the Nyquist frequency.

Because the folded spectrum is bandlimited in the same way as the spectrum of the input signal, the high-pass filter from Fig. 6.9 can be omitted. The combination of the folded signal with the input signal yields the most efficient method for an extension of the excitation signal towards high frequencies:

$$\begin{aligned}\tilde{u}_{\mathrm{wb}}(k) &= \tilde{u}_{\mathrm{nb}}(k) + \tilde{u}_{\mathrm{nb}}(k)\,\cos(\Omega_{\mathrm{M}}\,k)\\ &= \tilde{u}_{\mathrm{nb}}(k)\,(1+(-1)^k).\end{aligned} \tag{6.9}$$

The base-band of the input signal is preserved transparently if the upper band limit of the base-band speech is $\Omega_{\mathrm{bb,u}} < \pi/2$. Because of its efficiency, the method of spectral folding is used very frequently in bandwidth extension algorithms. Similar approaches can also be found in base-band speech codecs such as the GSM full-rate speech codec [69].

By the use of the spectral folding method, some systematic errors are produced, as can be observed in Fig. 6.10 (a). Since the fundamental frequency of the speech is not considered, the reproduced discrete structure in the extended frequency band is inconsistent during voiced sounds – the discrete frequency components are not correctly placed at integer multiples of the fundamental frequency, resulting in a metallic sound or 'ringing' of the enhanced speech $\tilde{s}_{\mathrm{wb}}(k)$. Further, the position of the extended frequency band is invariably determined by the sampling rate and the band limits of the input signal. In general, a spectral gap is created in the frequency range $\Omega_{\mathrm{bb,u}} < \Omega < \pi - \Omega_{\mathrm{bb,u}}$. For typical telephone speech (300 Hz to 3.4 kHz), for example, there will be a large gap between 3.4 and 4.6 kHz. In addition, the upper band limit of the folded signal is determined by the lower band limit of the input speech. For telephone speech, the upper limit of the extended speech is at 7.7 kHz. Serious artefacts are produced if there is a DC component in the input signal: the folded DC component yields a strong stationary sinusoid at the Nyquist frequency.

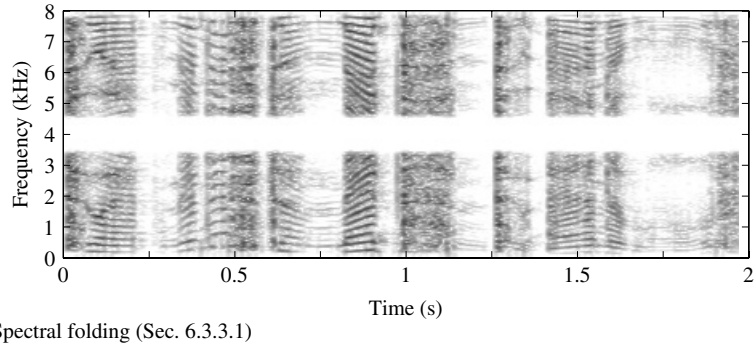
#### 6.3.3.2 Spectral Translation

In this section, the modulation shall be performed with a *fixed* modulation frequency as well. Now, the modulation frequency $\Omega_{\mathrm{M}}$ is specified by the bandwidth of the bandlimited input speech (lower band limit $\Omega_{\mathrm{bb,l}}$ and upper band limit $\Omega_{\mathrm{bb,u}}$)

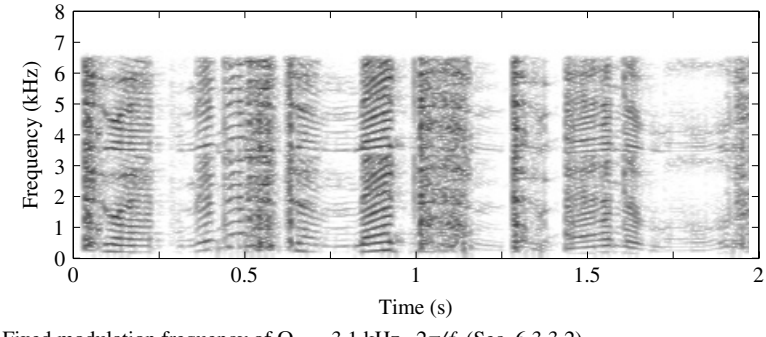
$$\Omega_{\mathrm{M}} = \Omega_{\mathrm{bb,u}} - \Omega_{\mathrm{bb,l}}. \tag{6.10}$$

Owing to this setting of $\Omega_{\mathrm{M}}$ the spectrum $\tilde{U}_{\mathrm{nb}}(\mathrm{e}^{j(\Omega-\Omega_{\mathrm{M}})})$, shifted towards high frequencies (see Eqn. 6.8), starts in continuation of the base-band spectrum $\tilde{U}_{\mathrm{nb}}(e^{j\Omega})$, that is, there is no gap in the spectrum of the extended speech. The upper band limit of the extended speech depends on the band limits of the base-band signal. It is defined by the frequency $\Omega_{\mathrm{bb,u}} + \Omega_{\mathrm{M}} = 2\,\Omega_{\mathrm{bb,u}} - \Omega_{\mathrm{bb,l}}$ (corresponding to 6.5 kHz for telephone speech).

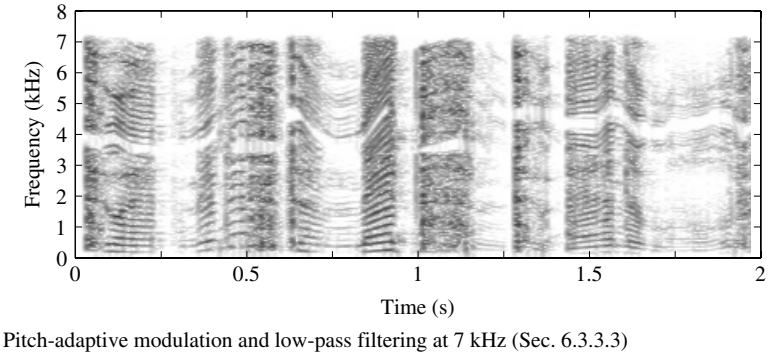
Prior to the mixing of the extended excitation with the base-band excitation signal, the frequency range in which the downwardly shifted spectrum is situated shall be removed by high-pass filtering (see Fig. 6.9). The cut-off frequency of the employed high-pass filter shall be equal to the upper band limit $\Omega_{\mathrm{bb,u}}$ of the base-band which is, due to the




(a) Spectral folding (Sec. 6.3.3.1)




(b) Fixed modulation frequency of $\Omega_M = 3.1\,\text{kHz} \cdot 2\pi/f_s$ (Sec. 6.3.3.2)




(c) Pitch-adaptive modulation and low-pass filtering at 7 kHz (Sec. 6.3.3.3)

**Figure 6.10** Spectrograms of the excitation signal $\tilde{u}_{wb}(k)$, extended via modulation-based techniques. Black regions reflect a large short-term power spectrum. The sentence 'to administer medicine to animals' is spoken by a female voice. Note that there is a pitch estimation error in the pitch-adaptive approach (c) after about 1.5 s

particular choice of the modulation frequency, also the lower limit of the spectrum shifted towards high frequencies. The desired stopband attenuation of the filter shall be obtained for frequencies below $\max\left(\Omega_{bb,l}, \Omega_{bb,u} - 2\,\Omega_{bb,l}\right)$, for example, below 2.8 kHz if typical telephone speech is the input of the BWE system.

#### 6.3.3.3 Pitch-adaptive Modulation

The modulation schemes with fixed modulation frequencies that have been described so far share the disadvantage that the discrete spectral structure of the extended excitation signal $\tilde{u}_{\mathrm{wb}}(k)$ during voiced sounds is inconsistent. To achieve a better performance, a further possibility to control the modulation frequency $\Omega_{\mathrm{M}}$ that takes the pitch frequency $\Omega_{\mathrm{p}}$ of the current speech frame into account shall be studied. The method has been developed independently by Fuemmeler and Hardie [77], Kornagel [147], and Jax and Vary [130], all in 2001. The idea to utilize information on the fundamental frequency of the speech for the extension of an excitation signal was first proposed by Makhoul and Berouti [167].

The basis of the *pitch-adaptive modulation* (PAM) in the time domain is an estimate $\tilde{\Omega}_{\mathrm{p}} = 2\pi\,\tilde{F}_0/f_{\mathrm{s}}$ of the fundamental frequency in the currently processed frame of the speech signal (e.g. Hess [109]). The modulation frequency $\Omega_{\mathrm{M}}$ is then adjusted in dependence of the estimate $\tilde{\Omega}_{\mathrm{p}}$, such that the shifted tonal components of the base-band excitation correspond to proper harmonics of the fundamental frequency within the extended frequency band

$$\Omega_{\mathrm{M}} = n_{\mathrm{M}}\,\tilde{\Omega}_{\mathrm{p}} \qquad \text{with } n_{\mathrm{M}} \in \mathbb{N}^{+} \text{ and } \Omega_{\mathrm{M,l}} \leq \Omega_{\mathrm{M}} \leq \Omega_{\mathrm{M,u}}. \tag{6.11}$$

In this way, for example, the $q$th harmonic of the fundamental frequency of the speech is shifted to the position of the $(q + n_{\mathrm{M}})$th harmonic. The integer-valued factor $n_{\mathrm{M}}$ is an adjustable parameter that has to be specified for each signal frame depending on the estimated fundamental frequency such that the resulting modulation frequency is between $\Omega_{\mathrm{M,l}}$ and $\Omega_{\mathrm{M,u}}$. By the limitation of the range of values of $\Omega_{\mathrm{M}}$, it shall be prevented that the bandwidth of the extended speech signal fluctuates strongly because of the variations of the fundamental frequency of the speech. The adaptive calculation of $n_{\mathrm{M}}(m)$ can, for example, be performed by the following method

$$n_{\mathrm{M}}(m) = \begin{cases} \left\lceil \frac{\Omega_{\mathrm{M,l}}}{\tilde{\Omega}_{\mathrm{p}}(m)} \right\rceil, & \text{if } n_{\mathrm{M}}(m-1)\,\tilde{\Omega}_{\mathrm{p}}(m) < \Omega_{\mathrm{M,l}} \\ n_{\mathrm{M}}(m-1), & \text{if } \Omega_{\mathrm{M,l}} \leq n_{\mathrm{M}}(m-1)\,\tilde{\Omega}_{\mathrm{p}}(m) \leq \Omega_{\mathrm{M,u}} \\ \left\lfloor \frac{\Omega_{\mathrm{M,u}}}{\tilde{\Omega}_{\mathrm{p}}(m)} \right\rfloor, & \text{if } n_{\mathrm{M}}(m-1)\,\tilde{\Omega}_{\mathrm{p}}(m) > \Omega_{\mathrm{M,u}}. \end{cases} \tag{6.12}$$

The basic principle of Eqn. 6.12 is to keep the factor $n_{\mathrm{M}}(m-1)$ that has been used in the preceding frame, if possible, thereby minimizing the number of switchings. If the reuse of the factor $n_{\mathrm{M}}(m)$ would lead to an under- or overshooting of the minimum or maximum modulation frequencies $\Omega_{\mathrm{M,l}}$ and $\Omega_{\mathrm{M,u}}$, respectively, the value of $n_{\mathrm{M}}(m)$ is corrected such that the new modulation frequency is just within the admissible range. The described procedure to control $n_{\mathrm{M}}(m)$ implies that the difference $\Omega_{\mathrm{M,u}} - \Omega_{\mathrm{M,l}}$ is greater than the maximum possible fundamental speech frequency such that there exists a valid factor $n_{\mathrm{M}}$, fulfilling the requirements from Eqn. 6.11, for each potential estimate $\tilde{\Omega}_{\mathrm{p}}$.

If the input signal of the bandwidth extension system has the typical telephone bandwidth (300 Hz to 3.4 kHz), the minimum modulation frequency should be specified by the bandwidth of the input speech $\Omega_{\mathrm{M,l}} = 3.1\,\mathrm{kHz} \cdot 2\pi/f_{\mathrm{s}}$. The upper limit $\Omega_{\mathrm{M,u}}$ of the

modulation frequency can, for example, be adjusted to $\Omega_{\mathrm{M,u}} = 4.6\,\mathrm{kHz} \cdot 2\pi/f_\mathrm{s}$ such that the maximum upper band limit of the extended speech is about 7 kHz. In the example of Fig. 6.10 (c), the maximum modulation frequency is set to $\Omega_{\mathrm{M,u}} = 3.6\,\mathrm{kHz} \cdot 2\pi/f_\mathrm{s}$, and the variations of the upper band limit of the extended excitation signal $\tilde{u}_\mathrm{wb}(k)$ are suppressed by low-pass filtering the modulated signal $\tilde{u}_\mathrm{M}(k)$ with a cut-off frequency of 7 kHz.

Since an absolutely accurate estimate of the fundamental frequency of the speech cannot by expected from any realizable pitch estimation algorithm, we have evaluated the impacts of typical estimation errors on the performance of the pitch-adaptive modulation method. A quite typical error of pitch estimation algorithms is *pitch doubling*, that is, $\tilde{\Omega}_\mathrm{p} = 2\,\Omega_\mathrm{p}$. Pitch-doubling errors, however, do not have any effect on the results of the pitch-adaptive modulation approach because the resulting modulation frequency $\Omega_\mathrm{M}$ is again an integer multiple of the true fundamental frequency $\Omega_\mathrm{p}$. In the case of relatively small deviations of the estimated pitch frequency $\tilde{\Omega}_\mathrm{p}$, on the other hand, the PAM algorithm exhibits a quite strong susceptibility because any such error is significantly increased by the multiplication with the large factor $n_\mathrm{M}$. Consequently, a very accurate pitch estimation algorithm is needed for the estimation of $\tilde{\Omega}_\mathrm{p}$. Otherwise, the positions of discrete tonal components in the extended frequency bands will be inconsistent, and the performances of the PAM algorithm and the fixed spectral translation (Sec. 6.3.3.2) will be to a certain extent alike. If, on the other hand, a sufficiently accurate pitch estimation algorithm is used with the pitch-adaptive modulation approach, the resulting speech signal $\tilde{s}_\mathrm{wb}(k)$ will sound more natural in comparison. The artefacts (metallic sound, 'ringing') that are produced by mis-aligned harmonics are reduced noticeably.

### *6.3.4 PITCH SCALING*

Finally, a new method for the extension of the excitation signal towards high frequencies that is based on frequency scaling of the original bandlimited speech signal shall be described. Figure 6.11 illustrates the concept of the approach by a block diagram. The application of the method for an exemplary speech signal is shown in the spectrograms in Fig. 6.12: first, by doubling the pitch frequency (pitch doubling), a version $\tilde{u}_\mathrm{eb}(k)$ of the excitation signal is produced, which has a doubled upper band limit in comparison to the bandlimited excitation signal $\tilde{u}_\mathrm{nb}(k)$. Comparing the spectrograms of the narrowband excitation $\tilde{u}_\mathrm{nb}(k)$ from Fig. 6.12 (a) and of the signal $\tilde{u}_\mathrm{eb}(k)$ after pitch

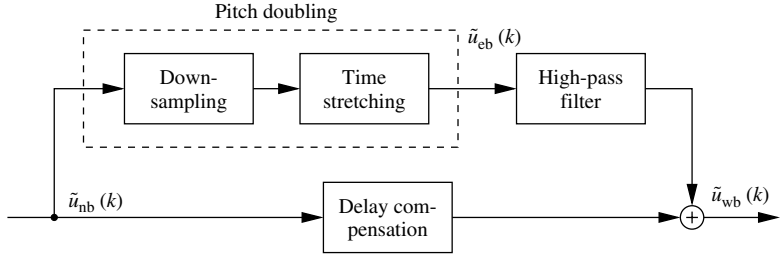
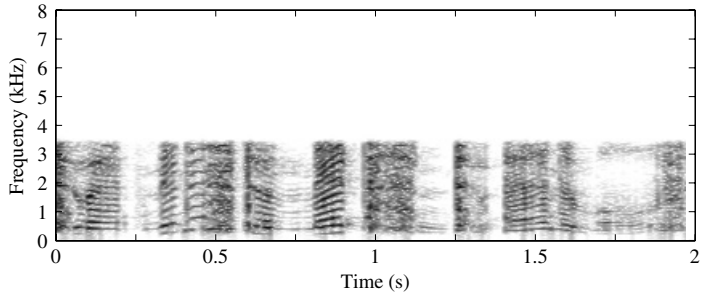
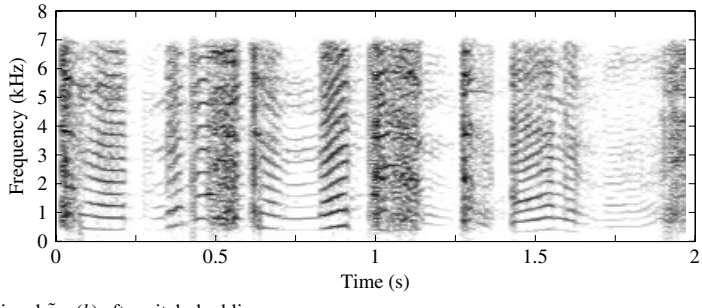


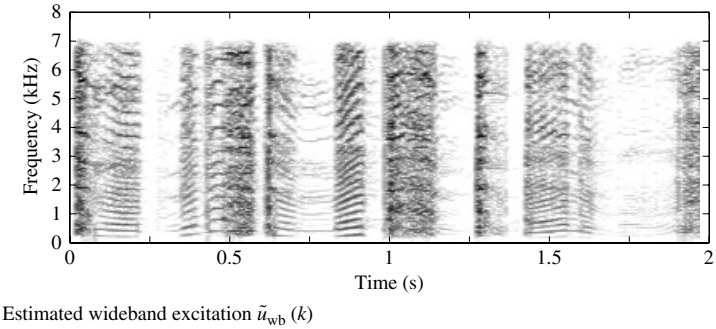

**Figure 6.11** Extension of the excitation signal via pitch scaling. The pitch doubling is realized by a downsampling by a factor of 2 and subsequent time stretching




(a) Estimated narrowband excitation signal $\tilde{u}_{\mathrm{nb}}(k)$




(b) Signal $\tilde{u}_{\mathrm{eb}}(k)$ after pitch doubling




(c) Estimated wideband excitation $\tilde{u}_{\mathrm{wb}}(k)$

**Figure 6.12** Spectrograms of intermediate signals for the pitch-scaling approach to the extension of the excitation signal. Black regions reflect a large short-term power spectrum. The sentence 'to administer medicine to animals' is spoken by a female voice

doubling in Fig. 6.12 (b), it can be observed that the latter signal has frequency components up to a cut-off frequency of about 6.8 kHz. During voiced sounds, the signal $\tilde{u}_{\mathrm{eb}}(k)$ contains tonal components only at *even* integer multiples of the fundamental frequency of the bandlimited excitation $\tilde{u}_{\mathrm{nb}}(k)$ from Fig. 6.12 (a).

The pitch doubling can, for example, be performed as depicted in Fig. 6.11: first, the sampling rate of the narrowband excitation $\tilde{u}_{\mathrm{nb}}(k)$ is reduced by a factor of 2[1],

[1] In the design of the downsampling method, an advantage can be taken here from the fact that the bandlimited excitation signal $\tilde{u}_{\mathrm{nb}}(k)$ already has an upper band limit that is lower than half of the Nyquist frequency. Therefore, a downsampling by a factor of 2 can simply be performed by omitting every other sample of the signal $\tilde{u}_{\mathrm{nb}}(k)$.

thereby effectively halving the length of signal segments (as expressed in a number of samples) In the second step, the downsampled signal is elongated conversely by employing time-scaling techniques (e.g. Verhelst and Roelands [289], Moulines and Verhelst [181], Verhelst [288]), that is, retaining the pitch frequency of the *downsampled* excitation signal. Since the impacts of the downsampling and time-scaling operations on the length of the speech segments as measured in samples per frame compensate each other, the serial concatenation of the two operations yields a doubling of the pitch of the signal components in $\tilde{u}_{\mathrm{nb}}(k)$.

After pitch doubling, the signal $\tilde{u}_{\mathrm{eb}}(k)$ is high-pass filtered and added to the properly delayed bandlimited excitation signal $\tilde{u}_{\mathrm{nb}}(k)$, thus yielding the estimated wideband excitation signal $\tilde{u}_{\mathrm{wb}}(k)$. It can be seen in Fig. 6.12 (c) that in those speech phases in which the bandlimited speech represents a harmonic complex tone, the pitch-scaling approach has the advantage that tonal signal components in the extended frequency band are at integer multiples of the fundamental frequency of the speech. Accordingly, there are no metallic or ringing artefacts in the enhanced speech. Note, however, that only harmonics with an even order are present in the extended band (above 3.4 kHz) due to the pitch doubling.

The method in general succeeds in regenerating the proper harmonic structure in the estimated wideband excitation signal. In some cases, particularly for speakers with a very low pitch frequency, the impression is produced by the algorithm that a second simultaneous speaker is present in the background of the enhanced speech who speaks with a doubled pitch in comparison to the original speaker. A particular advantage of this algorithm is the fact that it does not need any explicit estimate of any parameter, $F_0$, $V$, or $\sigma$, of the excitation part of the source model from Sec. 6.2.1. Accordingly, the method has a very high robustness.

### *6.3.5 DISCUSSION*

To evaluate the performances of the different methods for the extension of the excitation signal, extensive informal listening tests have been performed. To produce the speech samples for these tests, the set-up from Fig. 6.13 was used. The block diagram resembles our bandwidth extension system from Sec. 6.2.2, except that the estimation of the wideband spectral envelope from the upper signal path of Fig. 6.5 is replaced by an LP analysis of the *original* wideband speech. Thus, the AR coefficients $\mathbf{a}_{\mathrm{wb}}$ can be assumed

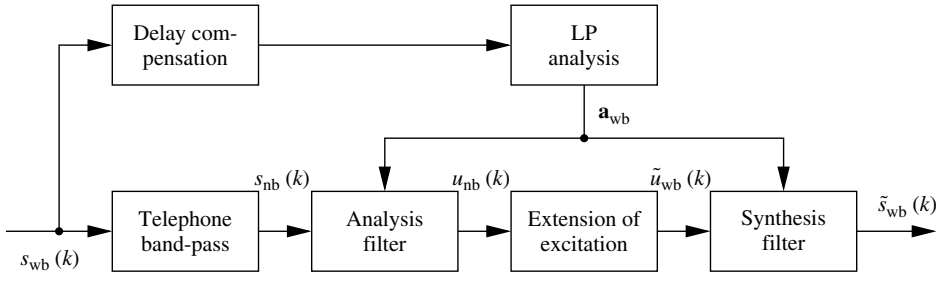



**Figure 6.13** System for the generation of speech samples for evaluating the quality of different approaches for the extension of the excitation signal via informal listening tests

to be optimal (in the sense of estimating the narrowband excitation signal $u_{\mathrm{nb}}(k)$), and potential artefacts in $\tilde{s}_{\mathrm{wb}}(k)$ are solely due to the extension of the excitation signal.

We have performed many informal listening tests that have shown that – on the precondition that the bandwidth extension of the spectral envelope works well – the human ear is amazingly insensitive to distortions of the excitation signal at high frequencies above 3.4 kHz. For example, spectral gaps of moderate width as produced by band-stop filters are almost inaudible. Further, inconsistencies of the harmonic structure of speech at high frequencies do not significantly degrade the subjective quality of the enhanced speech signal. The above comments particularly apply if the extended speech signal is played back via the physically constrained acoustical front-end of, for example, a mobile handset. As any such front-end in general has low-pass characteristics, the audibility of artefacts in the spectral fine structure of the enhanced speech $\tilde{s}_{\mathrm{wb}}(k)$ is reduced even further.

Owing to the beneficial properties of the human auditory system at high frequencies, all of the described methods for the extension of the excitation signal towards high frequencies perform well or very well if a good estimate of the wideband spectral envelope is available. A reasonable compromise between the maximization of the subjective quality of the output signal and the computational complexity is given by the modulation with the fixed modulation frequency of $\Omega_{\mathrm{M}} = \Omega_{\mathrm{bb,u}} - \Omega_{\mathrm{bb,l}}$.

The extension of the excitation signal towards low frequencies, on the other hand, is more difficult. The low-frequency components (e.g. below 300 Hz) are especially dominant during voiced sounds, and the human ear is rather sensitive to variations of the harmonic structure in this frequency range. Because most of the methods that are capable of extending the excitation towards low frequencies are based on a pitch estimation algorithm (with limited accuracy), the regenerated low-frequency harmonics often do not fit the harmonics within the base-band of the speech. This produces the distracting impression that a second simultaneous speaker is contained in the enhanced speech signal.

Nevertheless, the subjective quality of the speech signals generated by the system from Fig. 6.13, that is, with knowledge of the true spectral envelope of the wideband speech, is reasonably well, particularly for the extension towards high frequencies. This observation conforms to the results of previous investigations, where it was found that the quality of the estimated wideband spectral envelope is far more important for the subjective quality of the bandwidth-extended speech signal than the extension of the excitation signal (Carl [44]). This has also been recognized in high-frequency BWE for audio applications, see Sec. 5.5.

## 6.4 ESTIMATION OF THE WIDEBAND SPECTRAL ENVELOPE

The essential step in bandwidth extension algorithms is the estimation of the spectral envelope of the wideband speech signal. This task corresponds to the upper signal path in the block-diagram Fig. 6.5 of the bandwidth extension algorithm.

In most adaptive bandwidth extension algorithms, statistical estimation methods are used, which are to a certain extent similar to approaches from pattern recognition or speech recognition (see e.g. Fukunaga [79], Rabiner and Juang [216]). The estimation of the spectral envelope is in general performed in several consecutive steps as illustrated in Fig. 6.14. The three steps are executed for each frame of the speech signal and for each

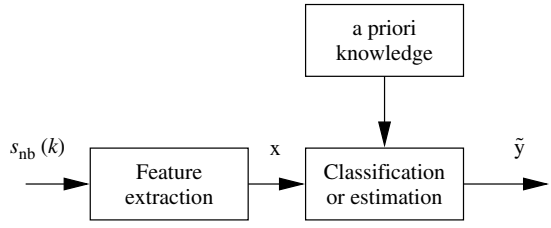



**Figure 6.14** Intermediate steps in the estimation of the representation $\tilde{\mathbf{y}}$ of the wideband spectral envelope

missing frequency band. Note that the frame index $m$ will be omitted in the following sections if it is not essential for the understanding of the particular topic.

***Feature extraction*** From each signal frame of the narrowband input speech $s_{nb}(k)$, several features $\mathbf{x}$ are extracted, which carry information on the state of the source model, that is, indirectly on the estimated spectral envelope of the missing frequency band. By the feature extraction algorithm, the dimension and complexity of the estimation problem are reduced significantly. The art is to find a *compact* set of features, that is, the dimension $\dim \mathbf{x}$ of the feature vectors $\mathbf{x}$ shall be low although the features shall carry much information to allow a proper estimation of $\mathbf{y}$. A more detailed description of the feature extraction step will be given in Sec. 6.5.

***A priori knowledge*** The extracted features $\mathbf{x}$ are compared with a priori knowledge, comprising information on the joint behaviour of the features $\mathbf{x}$ and the unknown quantity $\mathbf{y}$. Several representations are possible such as linked codebooks (tables), transformation matrices, reflecting linear correlation, or statistical models, for example, of the joint PDF $p(\mathbf{x}, \mathbf{y})$. The utilized representation of the a priori knowledge is strongly linked with the employed estimation method.

In general, the a priori knowledge has to be acquired during an off-line training phase before applying bandwidth extension. For this, a larger amount of wideband speech data is utilized. The model parameters will be stored for later use in the application phase of the BWE system.

***Classification or estimation*** The final step is the estimation of the representation $\mathbf{y}$ of the spectral envelope. This step can be based on different classification or estimation concepts, the most prominent of which are codebook mapping, linear or piecewise-linear mapping, and Bayesian estimation, according to the possible representations of the a priori knowledge as listed above. These approaches will be described together with the employed a priori knowledge in Secs. 6.6 to 6.9. In literature, sporadically also other approaches may be found such as, for example, neural networks (Tanaka and Hatazoe [262], Uncini *et al.* [277], Iser and Schmidt [119]). In some papers, fixed spectral envelopes are used (Larsen *et al.* [157]).

Note that commonly the required AR coefficients $\tilde{\mathbf{a}}$ of the wideband linear prediction filters are not estimated directly but some other mathematical representation of the

spectral envelope is determined first. To keep our presentation abstract, we will denote the estimator output by the quantity $\tilde{\mathbf{y}}$ in the sequel, assuming that $\tilde{\mathbf{y}}$ can directly be converted into the coefficient vector $\tilde{\mathbf{a}}$ to be applied in the analysis and synthesis filters from Fig. 6.5. Common representations will be described in Sec. 6.4.1.

To make things even more complicated, $\tilde{\mathbf{y}}$ may also stand for the shape and relative gain of the spectral envelope within a particular missing sub-band only. In this case, the wideband spectral envelope coefficients $\tilde{\mathbf{a}}$ have to be computed by evaluating $\tilde{\mathbf{y}}$ in addition to the available narrowband speech signal $s_{\mathrm{nb}}(k)$. This approach will be discussed in Sec. 6.4.1.1.

### *6.4.1 REPRESENTATIONS OF THE ESTIMATED SPECTRAL ENVELOPE*

It is an open question, what is the best representation $\mathbf{y}$ of the wideband spectral envelope to be estimated by the framework from Fig. 6.14. The spectral envelope can be represented in many different forms. From an algorithmic view, the most natural representation would be to directly use the AR coefficients $\mathbf{a}$ (i.e. $\tilde{\mathbf{y}} = \tilde{\mathbf{a}}$). In fact, this representation can be used if the estimation algorithm performs a hard classification, for example, by the codebook mapping approach (Carl [44], Carl and Heute [45], Yoshida and Abe [301]). Using just slightly more sophisticated estimation methods, for example, by averaging over the most likely codebook entries, there will be the problem that the stability of the LP synthesis filter $1/A(z)$ cannot be guaranteed. Therefore, direct estimation of AR coefficients is not often used for bandwidth extension.

The most frequently used representations of the spectral envelope in bandwidth extension of speech are line spectral frequencies (LSF), for example, in Enbom and Kleijn [64], Miet *et al.* [174], Chennoukh *et al.* [50]. They are defined as the roots of two symmetric and anti-symmetric polynomials reflecting the transfer function of the linear prediction filter as given in Eqn. 6.4 (Itakura [120]). The conversion from AR coefficients to LSF vectors and vice versa is unique. The outstanding advantage of the LSF representation is that there is a very simple rule to guarantee stable LP synthesis filters: the elements of the LSF vector have to be sorted in ascending order, and their values must be between zero and $\pi$. For detailed information on properties of LSF vectors, the reader is referred to the speech-coding literature, for example, Markel and Gray [168], Paliwal and Kleijn [197].

Other interesting representations of the wideband spectral envelope are cepstral coefficients. The real cepstrum of a signal frame is computed by an inverse discrete Fourier transform (DFT) of its logarithmized amplitude spectrum. Analogously, the log amplitude frequency characteristics of an AR filter can be approximated by a series of cepstral coefficients

$$\ln \frac{\sigma^2}{|A(\mathrm{e}^{j\Omega})|^2} = \sum_{i=-\infty}^{\infty} c_i \, \mathrm{e}^{-ji\Omega}, \tag{6.13}$$

where ln denotes the natural logarithm to the base of $e$, and $\sigma$ is a scalar gain factor. The cepstral coefficients $c_i$ are real valued and even ($c_{-i} = c_i$) owing to the minimum-phase frequency response of the all-pole filter $1/A(\mathrm{e}^{j\Omega})$ (Hagen [103]). The cepstral coefficients $c_0, c_1, \ldots$ can be calculated directly from the AR coefficients $\mathbf{a}$ and the gain factor $\sigma$ via

a simple recursive formula given, for example, in Markel and Gray [168]

$$c_0 = \ln \sigma^2$$

$$c_i = -a_i - \sum_{n=1}^{i-1} \frac{n}{i} c_n a_{i-n} \qquad \text{for } i > 0, \tag{6.14}$$

with $a_i = 0$ for $i > N_a$. Note that only the first $N_a + 1$ cepstral coefficients derived in this way are non-redundant, while the remaining ones can be determined from these foremost coefficients.

A particular advantage of the cepstral representation of AR coefficients is that a minimum mean square error (MMSE) solution for the estimated coefficients $\tilde{\mathbf{y}} = \tilde{\mathbf{c}}$ corresponds to a minimization of the log spectral distortion (LSD) measure

$$\begin{aligned} d_{\mathrm{LSD}}^2 &= \left(\frac{10}{\ln 10}\right)^2 \sum_{i=-\infty}^{\infty} (c_i - \tilde{c}_i)^2 \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(20 \log_{10} \frac{\sigma}{|A(\mathrm{e}^{j\Omega})|} - 20 \log_{10} \frac{\tilde{\sigma}}{|\tilde{A}(\mathrm{e}^{j\Omega})|}\right)^2 \mathrm{d}\Omega. \end{aligned} \tag{6.15}$$

The interest in this distortion measure is motivated by the fact that in speech coding the log spectral distortion correlates reasonably well with the subjective speech quality. Therefore, it has found wide acceptance in speech coding to assess the quality of quantizers of representations (i.e. parameters/coefficients) of the spectral envelope. Cepstral coefficients have been used for bandwidth extension of speech, for example, in Avendano *et al.* [23], Park and Kim [200], and Nilsson and Kleijn [187].

#### 6.4.1.1 Sub-band-based Assembly of the Wideband Spectral Envelope

If we consider the estimation of the wideband spectral envelope, we must distinguish between missing and available sub-bands: information on the spectral envelope within the frequency range of the narrowband input speech $s_{\mathrm{nb}}(k)$ of the BWE algorithm can be determined by conventional linear prediction techniques. For the missing frequency band(s), on the other hand, more or less sophisticated estimation methods as illustrated in Fig. 6.14 have to be used.

This distinction leads to the concept of subdividing the estimation of the (single) wideband spectral envelope according to the different frequency bands in the wideband speech (Jax [128]). The approach is illustrated in Fig. 6.15. First, the estimation of the shape and (relative) gain of the spectral envelope of the missing frequency band(s) is performed individually, according to Fig. 6.14. Each estimator of an individual sub-band spectral envelope can be tuned optimally to the specific properties of the particular frequency band. The separate estimators output their results in a short-term power spectrum domain. Then, the assembly of a joint description for the full frequency range of the wideband speech signal can be performed by simple concatenation of the estimated power spectra from the

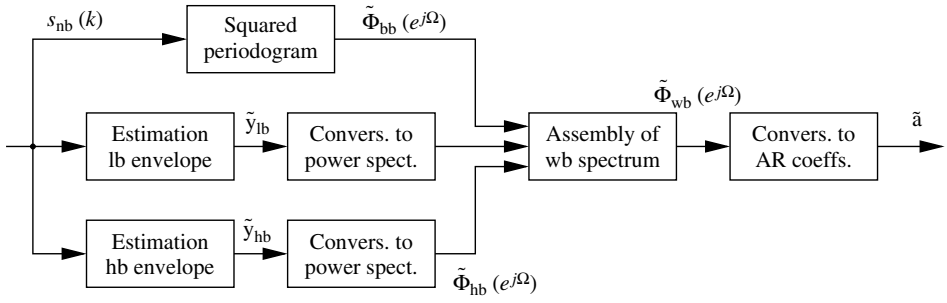



**Figure 6.15** Block diagram of the sub-band approach to estimate the AR coefficient set $\tilde{\mathbf{a}}$ of the wideband speech signal. The sub-band estimators for each missing frequency band consist of individual statistical estimators as shown in Fig. 6.14

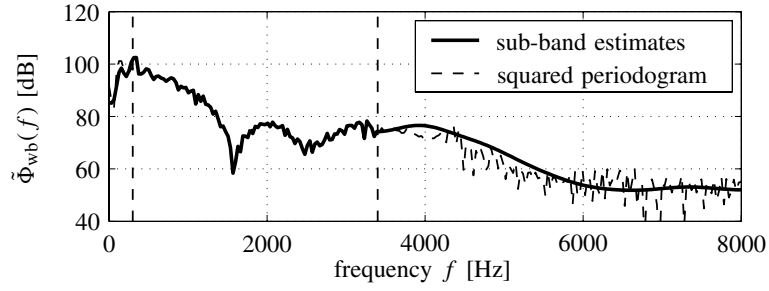
sub-bands. In a final step, the assembled smoothed power spectrum of the wideband signal is converted into the corresponding AR coefficients $\tilde{\mathbf{a}}$ (e.g. applying an inverse DFT and the Levinson–Durbin algorithm), which can then be used in the wideband analysis and synthesis filters of the BWE algorithm.

In general, the aforementioned procedure can be performed with an arbitrary number of sub-bands. In Fig. 6.15, the algorithm is depicted for three sub-bands, that is, in addition to the base-band there are two missing frequency bands at low and high frequencies. Further details on the sub-band-based assembly of the wideband spectral envelope can be found in Jax [128], Jax and Vary [133].

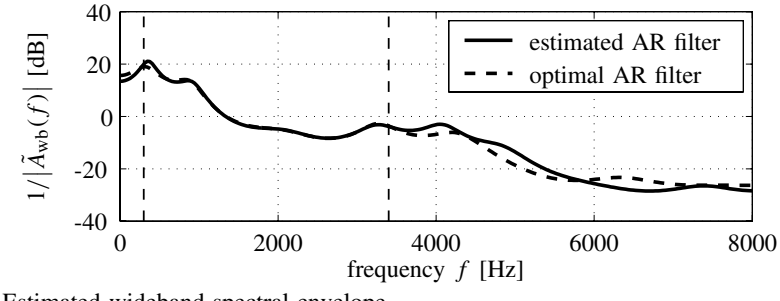
As illustrated in the lower diagram of Fig. 6.16 (b), the resulting AR coefficients constitute a good estimate of the wideband spectral envelope. In the base-band, the frequency response of the estimated AR filter (solid line) strongly matches the frequency response of the optimal AR filter as derived from the original wideband speech (dashed line). There are only slight deviations visible at the band edges due to errors in the estimated sub-band power spectra within the neighbouring missing frequency bands. In the extended frequency bands, there are some errors in the formant structure. Nevertheless, the ample run of the frequency response is estimated correctly. The common modelling of the spectral envelope of the base-band and extended bands has the advantage that the spectral envelope of the enhanced speech $\tilde{s}_{wb}(k)$ is smoothed in the transition regions between the bands.

### *6.4.2 INSTRUMENTAL PERFORMANCE MEASURE*

Before discussing various solutions for the sub-blocks of Fig. 6.14 in the remainder of this chapter, we will now define an instrumental performance measure needed to evaluate the different alternatives. Later, we are interested in what we can expect from bandwidth extension algorithms: will it be possible to perform as good as true wideband speech codecs by extending telephone speech or are there fundamental limits? To answer this question, BWE of speech shall be examined from an information theoretic perspective in Sec. 6.4.3.




(a) Concatenated short-term power spectrum




(b) Estimated wideband spectral envelope

**Figure 6.16** Example for the procedure of the sub-band-based estimation of the spectral envelope. In this example, there are two missing frequency bands, below 300 Hz and beyond 3.4 kHz. For comparison, the respective quantities as derived from the corresponding signal frame of the original wideband speech signal are shown by the dashed curves

#### 6.4.2.1 Auto-regressive Modelling of the Missing Sub-band Spectral Envelope

As described earlier, the spectral envelope of the enhanced speech signal is the principal key for a high subjective quality of the output signal of a bandwidth extension system. In Sec. 6.2.3, it was further pointed out that the bandlimited input signal $s_{\mathrm{nb}}(k)$ shall be contained transparently in the extended output signal $\tilde{s}_{\mathrm{wb}}(k)$. This can either be guaranteed implicitly by the structure of the algorithm as in our approach from Sec. 6.2.3, or the bandlimited input speech has to be considered explicitly as shown in Fig. 6.6. With both structures, there will only be errors within the spectral envelope of the extended frequency band of the output speech $\tilde{s}_{\mathrm{wb}}(k)$, and a performance measure should accordingly consider only distortions in this extended frequency band. Since a significant source of distortion is the attenuation or amplification of the spectral envelope in the extended frequency band with respect to the base-band spectral envelope, the *shape* as well as the *gain* of the spectral envelope of the extended band shall be investigated. The gain shall be expressed with respect to the base-band signal components, that is, it shall be a *relative* gain as specified later (compare Nilsson *et al.* [185], Park and Kim [200], Nilsson and Kleijn [187]).

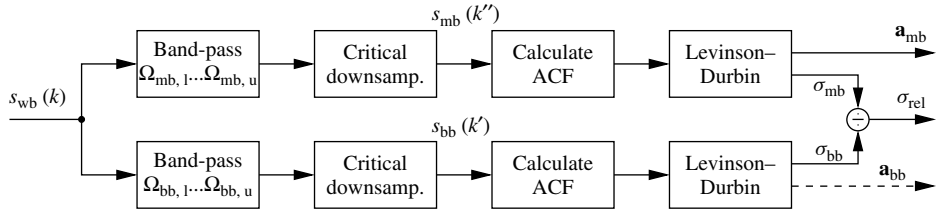



**Figure 6.17** Modelling of the spectral envelope of the missing (respectively extended) frequency band with respect to the base-band signal via two sub-band signals $s_{mb}(k'')$ and $s_{bb}(k')$ (non-general scheme: see footnote 3). For each sub-band signal, the parameters of an auto-regressive model are obtained from the estimated auto-correlation function (ACF) via a Levinson–Durbin recursion here. Note that the AR coefficient set $\mathbf{a}_{bb}$ of the base-band signal is not utilized

Without loss of generality, it shall be assumed in the following that the wideband speech signal is constituted from two sub-band signals: the base-band signal that corresponds to the input signal $s_{nb}(k)$ of the bandwidth extension system, and the sub-band signal containing the missing respectively extended frequency components[2]. The missing frequency band will be denoted by the subscript mb in the following text. It covers the frequencies between the lower band edge $\Omega_{mb,l}$ and the upper band edge $\Omega_{mb,u}$. The base-band of the speech signal starts at the lower cut-off frequency $\Omega_{bb,l}$ and ends at the upper cut-off frequency $\Omega_{bb,u}$ of the bandlimited input signal.

For the evaluation of an instrumental performance measure, it is presumed that the wideband speech signal $s_{wb}(k)$ is available. To define the spectral distortion measure, first the wideband speech signal is split into the two aforementioned sub-band signals (see Fig. 6.17). This is accomplished by band-pass filtering the wideband signal $s_{wb}(k)$, using two filters with lower and upper cut-off frequencies of $\Omega_{bb,l}$ and $\Omega_{bb,u}$, and $\Omega_{mb,l}$ and $\Omega_{mb,u}$, respectively. The band-pass-filtered signals are further critically downsampled such that the respective frequency components cover the whole frequency range of $\Omega = 0 \ldots \pi$ of the downsampled signals[3]. The two resulting sub-band signals are denoted by $s_{bb}(k')$ for the base-band signal, and by $s_{mb}(k'')$ for the signal containing the missing frequency band of the wideband speech. The sampling rates of the downsampled sub-band signals $s_{bb}(k')$ and $s_{mb}(k'')$ are

$$f_{s'} = f_s \frac{1}{\pi}(\Omega_{bb,u} - \Omega_{bb,l}) \quad \text{and} \quad f_{s''} = f_s \frac{1}{\pi}(\Omega_{mb,u} - \Omega_{mb,l}), \tag{6.16}$$

[2] Note that there are applications in which there are *two* (or even more) missing frequency bands, for example, at low (<300 Hz) and high (>3.4 kHz) frequencies. The performances of the two sub-band estimators can then be described individually using the approach of this chapter.

[3] Employing this procedure, it is inherently assumed that the ratios between the sampling rates of the wideband speech and of the downsampled sub-band signals are integer valued and that the band limits of the sub-bands are at suitable frequencies. In a more general scenario, an additional modulation of the sub-band signals is necessary to ensure that the band limits (e.g. $\Omega_{mb,l}$ and $\Omega_{mb,u}$) of the sub-bands in the wideband speech are mapped to the band limits 0 and $\pi$ of the critically downsampled signals.

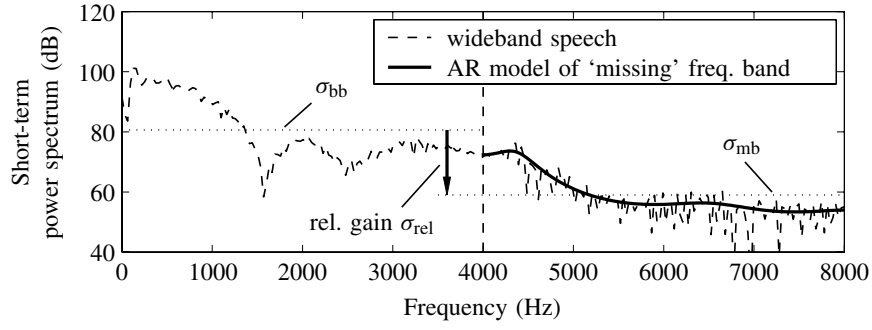



**Figure 6.18** Auto-regressive modelling of the spectral envelope of the sub-band signal containing the missing frequency band. In this example, the missing frequency band ranges from 4 to 8 kHz

respectively. The corresponding time indices or angular frequencies will be marked with one (for the base-band) or two (for the missing frequency band) apostrophes in the following.

In the next step, two individual auto-regressive models are fitted to frames of the two sub-band signals. For example, the model used for representing the missing frequency band spectrum $|S_{\mathrm{mb}}(\mathrm{e}^{j\Omega''})|^2$ is defined by (compare Fig. 6.18)

$$|S_{\mathrm{mb}}(\mathrm{e}^{j\Omega''})|^2 \approx \left|\frac{\sigma_{\mathrm{mb}}}{A_{\mathrm{mb}}(\mathrm{e}^{j\Omega''})}\right|^2 = \left|\frac{\sigma_{\mathrm{mb}}}{A_{\mathrm{mb}}(z'')}\right|^2_{z''=\mathrm{e}^{j\Omega''}} \tag{6.17}$$

with

$$A_{\mathrm{mb}}(z'') = 1 + \sum_{i=1}^{N_{a,\mathrm{mb}}} a_{\mathrm{mb},i}\,(z'')^{-i}. \tag{6.18}$$

The parameters of the models are estimated by conventional LP analysis, for example, by the Levinson–Durbin algorithm (Markel and Gray [168]). This is done individually for the two sub-band signals $s_{\mathrm{bb}}(k')$ and $s_{\mathrm{mb}}(k'')$. The results of the LP analysis are the coefficient set $\mathbf{a}_{\mathrm{mb}}$, representing the spectral envelope of the missing frequency band, as well as two gain factors $\sigma_{\mathrm{bb}}$ and $\sigma_{\mathrm{mb}}$ of the base-band and the missing frequency band, respectively. Since the *relative* gain of the extended frequency band shall be measured, we define $\sigma_{\mathrm{rel}} = \sigma_{\mathrm{mb}}/\sigma_{\mathrm{bb}}$. The order of $\mathbf{a}_{\mathrm{mb}} = \left[a_{\mathrm{mb},1}, a_{\mathrm{mb},2}, \ldots a_{\mathrm{mb},N_{a,\mathrm{mb}}}\right]^{\mathrm{T}}$ is $N_{a,\mathrm{mb}}$.

The parameters of the auto-regressive model of the missing frequency band of wideband speech that was defined in the previous paragraphs can alternatively be determined using *selective linear prediction* (SLP) techniques as described in Markel and Gray [168, Sec. 6.4]. Similar to the procedure described above (Fig. 6.17), the SLP approach allows to fit an auto-regressive model to a sub-band of the short-term spectrum of a signal. The resulting model corresponds to a critically sampled sub-band signal. Since the splitting of the sub-bands is performed in the frequency domain, the SLP method can flexibly be adapted to any bandwidth extension scenario (Jax [128]).

#### 6.4.2.2 Sub-band Log Spectral Distortion Measure

The performance of the estimation of the wideband spectral envelope shall be defined in terms of the *log spectral distortion* (LSD) of the missing frequency band. The squared LSD measure is specified in the frequency domain by (e.g. Markel and Gray [95], Gray, Buzo, Gray, and Matsuyama [96], Vary *et al.* [286])

$$d_{\mathrm{LSD}}^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(20\log_{10}\frac{\sigma_{\mathrm{rel}}}{|A_{\mathrm{mb}}(\mathrm{e}^{j\Omega''})|} - 20\log_{10}\frac{\tilde{\sigma}_{\mathrm{rel}}}{|\tilde{A}_{\mathrm{mb}}(\mathrm{e}^{j\Omega''})|}\right)^2 \mathrm{d}\Omega''. \tag{6.19}$$

Here, the quantities $A_{\mathrm{mb}}(\mathrm{e}^{j\Omega''})$ and $\sigma_{\mathrm{rel}}$ refer to the modelled frequency spectrum and relative gain of the missing frequency band of original wideband speech, and $\tilde{A}_{\mathrm{mb}}(\mathrm{e}^{j\Omega''})$ and $\tilde{\sigma}_{\mathrm{rel}}$ denote the corresponding estimated parameters as determined by a bandwidth extension system. Note that, because the LSD measure is evaluated for the critically downsampled sub-band signal $s_{\mathrm{mb}}(k'')$ containing only the missing frequency band, the integration range of $-\pi$ to $\pi$ in Eqn. 6.19 covers the missing frequency range in the original wideband speech signal. The unit of $d_{\mathrm{LSD}}$ is dB.

Unfortunately, the evaluation of the LSD measure in the frequency domain in general is quite complicated. Therefore, an alternative representation by a mean-square error criterion in the cepstral domain, following the definition from Eqn. 6.13, will be used in the following

$$\ln\frac{\sigma_{\mathrm{rel}}^2}{|A_{\mathrm{mb}}(\mathrm{e}^{j\Omega''})|^2} = \sum_{i=-\infty}^{\infty} c_i\,\mathrm{e}^{-ji\Omega''}. \tag{6.20}$$

With this definition, for a sequence of speech frames the *root mean square* (RMS) average of the LSD is given by

$$\overline{d}_{\mathrm{LSD}} = \frac{\sqrt{2}\,10}{\ln 10}\sqrt{E\left\{\frac{1}{2}(c_0-\tilde{c}_0)^2 + \sum_{i=1}^{\infty}(c_i-\tilde{c}_i)^2\right\}}. \tag{6.21}$$

Here, the function $E\{\cdot\}$ denotes the expectation operation.

Now, the output representation $\tilde{\mathbf{y}}$ of the estimation shall be defined in such a manner that the estimation performance can be determined by a mean-square error criterion. For this, the quantity $\mathbf{y}$ is defined as a *weighted* cepstral representation of the missing frequency band. It can be determined from the cepstral coefficients $c_0, c_1, \ldots$ that represent the AR model of the spectral envelope of the missing frequency band

$$y_i = \begin{cases} \frac{1}{\sqrt{2}}c_i, & \text{if } i = 0 \\ c_i, & \text{if } 1 \le i < \mathrm{d}. \end{cases} \tag{6.22}$$

The scalar values $y_i$ constitute the $d$-dimensional vector $\mathbf{y} = [y_0, y_1, \ldots y_{d-1}]^{\mathrm{T}}$. The dimension $d$ of $\mathbf{y}$ should be at least equal to $N_{a,\mathrm{mb}}$ such that all non-redundant cepstral

coefficients are considered. Inserting the definition of Eqn. 6.22 into Eqn. 6.21 yields the relationship

$$d_{\mathrm{LSD}}^2 \approx \left(\frac{\sqrt{2}\,10}{\ln 10}\right)^2 \sum_{i=0}^{d-1} (y_i - \tilde{y}_i)^2. \tag{6.23}$$

Note that the term on the right-hand side of Eqn. 6.23 is only an approximation of the log spectral distortion (Eqn. 6.19) because only the first $d$ summands of the sum in Eqn. 6.21 are considered.

To get a notion of the admissible sub-band log spectral distortion in bandwidth extension, the LSD performances of several wideband speech codecs (G.722 and AMR-WB at several bit rates) were investigated in Jax [128]. It was found that the wideband codecs achieve a near-transparent subjective speech quality even with an RMS LSD of more than 2 dB for the low-frequency band from 50 to 300 Hz, and with an RMS LSD of more than 3 dB for the high-frequency band from 3.4 to 7 kHz. Thus, it is conjectured that it is also possible in the BWE application to obtain a 'near-transparent' speech quality for RMS LSD values about 2 to 3 dB.

### *6.4.3 THEORETICAL PERFORMANCE BOUND*

It is plausible that an extension of the bandwidth of speech signals is only possible, if there are sufficient dependencies between the available bandlimited speech signal and the missing frequency components. The fact that the narrowband speech and the missing signal components are results of the same physical speech production process gives rise to the assumption that there are such dependencies in speech signals. This assumption is supported by the success of many BWE methods published throughout the last two decades. There are only few publications, however, that shade some light onto the information theoretic background of artificial bandwidth extension (Nilsson *et al.* [185, 186], Nordén *et al.* [188], Jax and Vary [131], Yang *et al.* [297], Epps [65]).

In digital signal processing, *linear* dependencies between signals are commonly described in terms of *correlation* factors. In an information theoretic perspective, the dependencies between different signals are described by their mutual information (MI), for example, Cover and Thomas [52]. In contrast to the correlation measure, mutual information covers all kinds of linear and non-linear dependencies. The aim of this section is to investigate the relationship between an upper bound on the achievable quality of a BWE algorithm (measured in terms of the instrumental performance measure from the previous section) on one hand, and the mutual information between representations of the bandlimited speech (feature vector $\mathbf{x}$) and of the missing frequency components (sub-band spectral envelope $\mathbf{y}$) on the other hand.

The quantity $\tilde{\mathbf{y}}$ as defined in Eqn. 6.22 can be calculated for *any* BWE algorithm, either from the extended speech signal or directly from the estimated representation of the wideband spectral envelope (e.g. as $\tilde{\mathbf{a}}$ in Fig. 6.5) if applicable. Therefore, an upper bound on the performance of a generalized BWE algorithm – measured in terms of the RMS LSD or the mean-square error of $\tilde{\mathbf{y}}$, respectively – also constitutes a bound on the performance of any other BWE algorithm with a differing representation of the wideband spectral envelope.

If we assume a memoryless, deterministic estimation[4] of $\tilde{\mathbf{y}} = f(\mathbf{x})$ of the missing spectral envelope from the feature vector $\mathbf{x}$, a relation between the mutual information $I(\mathbf{x};\mathbf{y})$, expressed in *nats* (Cover and Thomas [52]), and the minimum possible mean square estimation error of $\tilde{\mathbf{y}}$ can be formulated (Jax and Vary [131], Jax [128])

$$E\{\|\mathbf{y}-\tilde{\mathbf{y}}\|^2\} \geq \frac{\mathrm{d}}{2\pi\mathrm{e}} \exp\left(\frac{2}{\mathrm{d}}\big(h(\mathbf{y}) - I(\mathbf{x};\mathbf{y})\big)\right). \tag{6.24}$$

The relation depends on the differential entropy $h(\mathbf{y}) = -E\big\{\ln p(\mathbf{y})\big\}$, which comprises statistical properties of the estimated quantity $\mathbf{y}$. Further, $d = \dim \mathbf{y}$.

Owing to the weighting of the representation $\mathbf{y}$ of the missing frequency band, the mean-square error $E\{\|\mathbf{y}-\tilde{\mathbf{y}}\|^2\}$ resembles a truncated version of the cepstral distance within the square root of Eqn. 6.21. Because the truncated elements are non-negative, we find the inequality

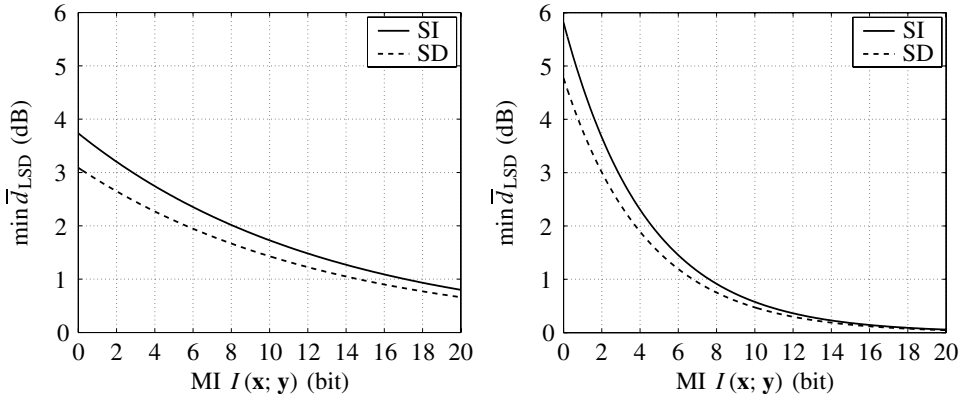
$$\begin{aligned}\overline{d}_{\mathrm{LSD}} &\geq \frac{\sqrt{2}\,10}{\log 10}\sqrt{E\big\{\|\mathbf{y}-\tilde{\mathbf{y}}\|^2\big\}} \\ &\geq \frac{\sqrt{2}\,10}{\log 10}\sqrt{\frac{\mathrm{d}}{2\pi\mathrm{e}}} \exp\left(\frac{1}{\mathrm{d}}\big(h(\mathbf{y}) - I(\mathbf{x};\mathbf{y})\big)\right).\end{aligned} \tag{6.25}$$

which gives a lower bound on the achievable RMS log spectral distortion in dependence of the mutual information $I(\mathbf{x};\mathbf{y})$ and differential entropy $h(\mathbf{y})$. While the differential entropy $h(\mathbf{y})$ only depends on statistical properties of the cepstral representation $\mathbf{y}$, the mutual information $I(\mathbf{x};\mathbf{y})$ additionally depends on the chosen feature set $\mathbf{x}$.

In the following, we will give examples of bounds defined by Eqn. 6.25 for the bandwidth extension of telephone speech. Two applications will be considered: extension towards low frequencies below 300 Hz and extension to high frequencies above 3.4 up to 7 kHz. In Jax [128], differential entropies $h(\mathbf{y})$ have been approximated for these applications by numerical simulations using the large SI100 speech corpus from the *Bayerisches Archiv für Sprachsignale* (BAS) (Schiel [234]). The SI100 corpus contains more than 35 hours of continuous wideband German speech spoken by 101 male and female speakers. The resulting bounds for the two aforementioned applications are illustrated in Fig. 6.19 in dependence of the mutual information. Similar investigations have been reported for the TIMIT database in Jax and Vary [131]. Details on the mutual information obtained for different feature sets will be presented in Sec. 6.5.

Unfortunately, it is in the nature of the information theoretic bound that it does not point out a particular strategy to design estimators that are optimal in the sense of minimizing the log spectral distortion measure. However, from the dependency of the performance bounds from the mutual information $I(\mathbf{x};\mathbf{y})$, it can be concluded that it is advantageous to select the elements of the utilized feature vector $\mathbf{x}$ such as to maximize the mutual information $I(\mathbf{x};\mathbf{y})$ (compare Sec. 6.5). To obtain the best possible quality with the BWE

[4] Note that we do not assume anything about the particular realization of the estimator: the function $f(\mathbf{x})$ can be interpreted as a generalized description of any kind of linear or non-linear classification or estimation with arbitrary complexity–including the approaches commonly used for BWE, for example, codebook mapping, linear mapping, statistical estimation, and neural networks.




(a) Towards high frequencies (3.4–7 kHz) (b) Towards low frequencies (0–300 Hz)

**Figure 6.19** Lower bounds on the RMS log spectral distortion $\overline{d}_{\mathrm{LSD}}$ for memoryless bandwidth extension of telephone speech (300 Hz–3.4 kHz). The bounds are given for speaker-independent (SI) and speaker-dependent (SD) solutions

system, it appears favourable – at least if the available mutual information $I(\mathbf{x};\mathbf{y})$ is low – to aim at speaker-dependent approaches.

Since the theoretical bound of Eqn. 6.25 is not tight, however, high mutual information $I(\mathbf{x};\mathbf{y})$ only is a necessary but not a sufficient condition for achieving a high performance with a specific feature vector $\mathbf{x}$. Thus, by observing the bound in Fig. 6.19, we cannot acquire knowledge on what we can expect from BWE algorithms but only on what we can *not* expect. We will see at the end of Sec. 6.5 that by selection of proper elements to include in the feature vector $\mathbf{x}$ it was to date only possible to achieve a mutual information of about 3 bit at best. Inserting this value into Eqn. 6.25 results in lower bounds of about 3 dB for speaker-independent solutions in both considered applications. Consequently, we can safely conclude that it is not possible, at least with the investigated features, to outperform the quality of wideband speech codecs by bandwidth extension.

## 6.5 FEATURE SELECTION

In this section, the focus shall be on the feature extraction block that is preceding the step of estimating the wideband spectral envelope (see Fig. 6.14). In general, feature extraction and estimation of the wideband spectral envelope are performed on a frame-by-frame basis with frame lengths of about 10 to 30 ms. The feature extraction reduces the dimensionality of each frame of the narrowband signal $s_{\mathrm{nb}}(k)$ such that the subsequent estimation of the spectral envelope representation is feasible and computationally efficient. The result is the feature vector $\mathbf{x} = [x_1, \ldots x_b]^{\mathrm{T}}$ with the dimension $b = \dim \mathbf{x}$. Usually, representations of the spectral envelope of the narrowband signal $s_{\mathrm{nb}}(k)$ are used as features, for example, LPC or LSF vectors or cepstral coefficients. In some contributions, additional features such as voicing criteria are taken into account. Here, we want to find some measures that help in finding the best composition of the feature vector $\mathbf{x}$.

The optimal feature extraction method (in the sense of high-quality bandwidth extension) for a fixed dimension of the feature vector allows the BWE algorithm to achieve the best subjective performance as compared to all other possible mappings with the same dimension. Unfortunately, evaluation and comparison of the subjective performances for a large number of alternative algorithms and/or feature sets $\mathbf{x}$ is very time consuming. Therefore, other means have to be used to assess the 'quality' of single features or feature vectors – instrumental measures are needed that provide suggestive hints for the selection of the best feature set.

In the next two sub-sections, two instrumental measures from information theory and statistics will be reviewed. In Sec. 6.5.3, the *linear discriminant analysis* (LDA) that results from an optimization of the separability measure will be introduced. With an LDA, the dimension of a feature vector can be reduced while the maximum discriminating power of the features is retained. In the last sub-sections, the usability of different features, well-tried and new ones, for the BWE problem will be evaluated using the introduced measures and procedures. The insights and results of this section are mostly independent from the particular approach used for estimating the wideband spectral envelope.

### *6.5.1 MUTUAL INFORMATION*

Shannon's *mutual information* (MI) $I(\mathbf{x};\mathbf{y})$ gives the mean information we gain on the estimated wideband spectral envelope representation $\mathbf{y}$ by knowledge of the feature vector $\mathbf{x}$. Mutual information can be regarded as an indication of the feasibility of the estimation task (Nilsson *et al.* [186], Jax and Vary [131]): in Sec. 6.4.3, it has been shown that for a specific mutual information $I(\mathbf{x};\mathbf{y})$ the minimum achievable mean-square estimation error $E\left\{\|\mathbf{y}-\tilde{\mathbf{y}}\|^2\right\}$ is lower bounded. The larger the MI, the lower is the bound. Hence, a large mutual information $I(\mathbf{x};\mathbf{y})$ is a *necessary* condition for high-quality estimation of $\mathbf{y}$ from the observations $\mathbf{x}$. Now, we want to investigate the mutual information for different features $\mathbf{x}$ of the narrowband speech signal.

For estimating the mutual information $I(\mathbf{x};\mathbf{y})$, we have to use a parametric approach because of the high dimension of the continuous vectors $\mathbf{x}$ and $\mathbf{y}$. The joint probability density function (PDF) $p(\mathbf{x},\mathbf{y})$ is approximated by a Gaussian mixture model (GMM) $\tilde{p}(\mathbf{x},\mathbf{y})$, that is, a sum of $L$ weighted multivariate Gaussian densities $\mathcal{N}(\cdot)$ with mean vectors $\mu_l$ and covariance matrices $\mathbf{V}_l$

$$\tilde{p}(\mathbf{x},\mathbf{y}) = \sum_{l=1}^{L} \rho_l \, \mathcal{N}(\mathbf{x},\mathbf{y};\mu_l,\mathbf{V}_l) \approx p(\mathbf{x},\mathbf{y}). \tag{6.26}$$

The scalar weights $\rho_l$ and the parameters $\mu_l$ and $\mathbf{V}_l$ of the individual Gaussians are trained by the expectation-maximization (EM) algorithm[5]. Then, the mutual information is estimated numerically from the parameters of the GMM (Hedelin and Skoglund [106])

$$I(\mathbf{x};\mathbf{y}) \approx E_{\tilde{p}(\mathbf{x},\mathbf{y})}\left\{\log \frac{\tilde{p}(\dot{\mathbf{x}},\dot{\mathbf{y}})}{\tilde{p}(\dot{\mathbf{x}})\,\tilde{p}(\dot{\mathbf{y}})}\right\}$$

[5] Further details on Gaussian mixture models can be found in Sec. 6.8.

$$\approx \frac{1}{M} \sum_{\nu=1}^{M} \log \frac{\tilde{p}(\dot{\mathbf{x}}(\nu), \dot{\mathbf{y}}(\nu))}{\tilde{p}(\dot{\mathbf{x}}(\nu))\, \tilde{p}(\dot{\mathbf{y}}(\nu))}. \tag{6.27}$$

In the computation of Eqn. 6.27, the vector pairs $\dot{\mathbf{x}}(\nu), \dot{\mathbf{y}}(\nu)$ are generated synthetically according to the model PDF $\tilde{p}(\mathbf{x}, \mathbf{y})$. In our investigations presented in Sec. 6.5.5, we have used $L = 256$ Gaussians with full covariance matrices. The numerical evaluation of Eqn. 6.27 was performed with $M = 10^6$ synthetic vector pairs (Jax [128]).

From the definition of mutual information, for example, Cover and Thomas [52], the following important properties of this measure for feature selection can be found:

- If the relation between two different feature vectors is defined by a bijective mapping, the MI is identical for both feature vectors. In this case, the MI measure does not provide any hint on which feature set shall be preferred.
- If several parameters of the narrowband speech (say $x_A$, $x_B$ and $x_C$) form a Markov chain $x_A \rightarrow x_B \rightarrow x_C$, that is, if $x_C$ is calculated from $x_B$, and $x_B$ is calculated from $x_A$, it appears favourable to select the very first element $x_A$ of the chain as a feature. Owing to the data-processing inequality (Cover and Thomas [52]), MI is maximized by this choice.
- For combined feature vectors, the MI cannot be simply added. In general, the MI has to be estimated again for the new vector.

### *6.5.2 SEPARABILITY*

From the field of pattern recognition, the *separability* is known as a measure of the quality of a particular feature set for a classification problem (Fukunaga [79]). In the BWE application, the class definitions should best be adopted to the method used to estimate the wideband spectral envelope: for example, if codebook mapping is used (Carl [44]), the classes should correspond to the correct codebook indices as computed from true wideband speech. For an HMM-based approach (Jax and Vary [133]), the classes should be the true HMM state information.

The separability measure can be calculated from a labelled set of training data, that is, for each feature vector in the set the corresponding class must be known. Let $\Xi_i$ denote the set of feature vectors $\mathbf{x}$ assigned to the $i$th class. The number of feature vectors in the $i$th set is $N_{\Xi_i} = |\Xi_i|$. The constant $N_{\mathcal{S}}$ denotes the number of classes. Then, the total number of frames in the training data is given by $N_m = \sum_{i=1}^{N_{\mathcal{S}}} N_{\Xi_i}$. From the labelled training data, the *within-class* covariance matrix

$$\mathbf{V}_{\mathbf{x}} = \frac{1}{N_m} \sum_{i=1}^{N_{\mathcal{S}}} \sum_{\mathbf{x} \in \Xi_i} (\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^T \tag{6.28}$$

and the *between-class* covariance matrix

$$\mathbf{B}_{\mathbf{x}} = \sum_{i=1}^{N_{\mathcal{S}}} \frac{N_{\Xi_i}}{N_m} (\mu_i - \mu)(\mu_i - \mu)^T \tag{6.29}$$

are calculated, where

$$\mu_i = \frac{1}{N_{\Xi_i}} \sum_{\mathbf{x} \in \Xi_i} \mathbf{x} \quad \text{and} \quad \mu = \sum_{i=1}^{N_S} \frac{N_{\Xi_i}}{N_m} \mu_i . \tag{6.30}$$

The separability measure shall be larger if the between-class covariance gets larger and/or if the within-class covariance gets smaller. Accordingly, the separability measure is empirically defined by the term $\mathbf{J}_\mathbf{x} = \mathbf{V}_\mathbf{x}^{-1} \mathbf{B}_\mathbf{x}$. To obtain a scalar measure for the separability of the classes, a trace criterion is used (Fukunaga [79])

$$\zeta(\mathbf{x}) = \operatorname{tr} \mathbf{J}_\mathbf{x} = \operatorname{tr} \left( \mathbf{V}_\mathbf{x}^{-1} \mathbf{B}_\mathbf{x} \right). \tag{6.31}$$

The separability depends on the definition of the classes. Comparing $\zeta(\mathbf{x})$ for different feature vectors $\mathbf{x}$ with the same class definitions, a larger value indicates a better suitability of the corresponding feature vector for classification and estimation.

The separability measure has the following properties:

- The definition of the separability measure is based on the implicit assumption of a normal distribution of the feature vectors that are assigned to each class. If this assumption is not valid, the significance of the separability measure is reduced.
- By the separability measure, all classes are treated alike. Therefore, the separability of two very similar classes (w.r.t. the represented speech sound) is rated like the separability of two very different classes. Hence, maximizing the separability does not necessarily lead to the optimum achievable estimation performance (e.g. in the MMSE sense) of the subsequent estimation rule.
- In general, the values of the separabilities cannot be added up if several features are assembled to a composite feature vector. In this case, the separability of the composite feature vector must be measured anew.

### *6.5.3 LINEAR DISCRIMINANT ANALYSIS*

The purpose of the *linear discriminant analysis* (LDA) is to obtain a feature vector with maximal compactness (Fukunaga [79]): starting from a high-dimensional 'super-vector' $\mathbf{x}_0$, the dimension of the feature vector $\mathbf{x}$ shall be reduced, while the discriminating power shall be retained or decreased as little as possible. The reduction of dimension is performed (in the BWE application phase) by means of a linear transformation

$$\mathbf{x} = \mathbf{H}^\mathrm{T} \mathbf{x}_0, \tag{6.32}$$

where the matrix $\mathbf{H}$ is a $\beta \times b$ matrix with $b = \dim \mathbf{x} < \beta = \dim \mathbf{x}_0$. The column vectors in $\mathbf{H}$ shall be linearly independent.

The matrix $\mathbf{H}$ is optimized such that the separability of $\mathbf{x}$ is maximized (Fukunaga [79])

$$\mathbf{H} = \arg \max_\mathbf{H} \zeta(\mathbf{x}), \quad \text{where} \tag{6.33}$$

$$\zeta(\mathbf{x}) = \operatorname{tr} \left( \mathbf{V}_\mathbf{x}^{-1} \mathbf{B}_\mathbf{x} \right) = \operatorname{tr} \left( \mathbf{H}^\mathrm{T} \mathbf{V}_{\mathbf{x}_0}^{-1} \mathbf{B}_{\mathbf{x}_0} \mathbf{H} \right).$$

The solution to Eqn. 6.33 is achieved by composing the matrix $\mathbf{H}$ from the eigenvectors $\Phi_1, \Phi_2 \ldots \Phi_b$ that are assigned to the $b$ largest eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_b$ of $\mathbf{V}_{\mathbf{x}_0}^{-1}\mathbf{B}_{\mathbf{x}_0}$. The computationally complex preparation of the transformation matrix $\mathbf{H}$ is performed only once, off-line during the training phase of the BWE algorithm.

The LDA makes it possible to take many primary features of the bandlimited speech signal into account, using a high-dimensional super-vector $\mathbf{x}_0$. Nevertheless, the dimension of $\mathbf{x}$ can be small – without loosing too much discriminating power – such that the computational complexity and memory consumption of the subsequent estimation algorithm are low.

### *6.5.4 PRIMARY FEATURES*

In the following text, brief definitions and descriptions of features of the narrowband speech $s_{\mathrm{nb}}(k)$ typically used in bandwidth extension algorithms for speech are given. For further particulars on specific features, the reader is referred to Jax [128] and the cited literature.

***Coefficients of the auto-correlation function (ACF)*** are often used for the voiced/unvoiced classification of speech segments (Campbell and Tremain [43], Wang [294]). Especially, the normalized first coefficient of the ACF is used because it reflects the spectral tilt of the signal spectrum. The normalized coefficients of the ACF can be estimated as follows

$$x_{\mathrm{acf}}(\lambda) = \frac{\sum_{\kappa=\lambda}^{N_\kappa - 1} s_{\mathrm{nb}}(\kappa - \lambda)\, s_{\mathrm{nb}}(\kappa)}{\sum_{\kappa=0}^{N_\kappa - 1} \left(s_{\mathrm{nb}}(\kappa)\right)^2}, \tag{6.34}$$

where $\lambda$ denotes the index of the ACF coefficient, and $N_\kappa$ is the number of samples per frame. Later, we will investigate the first ten coefficients ($\lambda = 1 \ldots 10$) and the ACF at the pitch lag (see below).

***Coefficients of a linear prediction filter (LPC)*** In *linear predictive coding* (LPC) of speech signals, FIR prediction filters with the coefficients $a_i, i = 1 \ldots N_a$ are used (compare Sec. 6.2.1). The prediction filter is described by the difference equation

$$\tilde{s}_{\mathrm{nb}}(k) = -\sum_{i=1}^{N_a} a_i\, s_{\mathrm{nb}}(k - i). \tag{6.35}$$

The optimal (in the sense of minimizing the power of the error signal $s_{\mathrm{nb}}(k) - \tilde{s}_{\mathrm{nb}}(k)$) filter coefficients $a_i$ are derived from the first coefficients of the auto-correlation function of the speech signal $s_{\mathrm{nb}}(k)$ utilizing, for example, the Levinson–Durbin algorithm (Makhoul [166], Markel and Gray [168]).

***The line spectral frequencies (LSFs)*** are an alternative representation of the LPC coefficients from the previous paragraph (Itakura [120]). The LSF coefficients have several advantageous properties with regard to coding and interpolation (Paliwal [196], Vary *et al.* [286]), compare Sec. 6.4.1. Therefore, this representation of the LPC coefficients is often used in speech codecs. It has been used for bandwidth extension in Miet *et al.* [174], Chennoukh *et al.* [50].

***LPC-derived cepstral coefficients (LPC-cepstrum)*** The transfer function of an LP synthesis filter can be represented by an infinite sequence of cepstral coefficients (e.g. Sec. 6.4.1, Markel and Gray [168], Hagen [103]). The according cepstral coefficients $c_i$ are calculated from the prediction coefficients $a_i$ by the simple recursive formula from Eqn. 6.14. The cepstral representation has the advantage of a good decorrelation of the coefficients. This is advantageous for modelling the PDF $p(\mathbf{x})$. LPC-derived cepstral coefficients have been used for BWE first in Abe and Yoshida [11] and Avendano *et al.* [23].

***Linear cepstral coefficients*** A cepstral representation of the spectral envelope of the speech signal can alternatively be calculated from the magnitude spectrum of the signal frame. For this purpose, the speech frame is transformed into the frequency domain via a *discrete Fourier transform* (DFT). Then, the logarithm is applied to the magnitude spectrum, and the result is transformed to the cepstral domain with an inverse DFT (Oppenheim and Schafer [194]). If the cepstrum is truncated, the coefficients represent a cepstrally smoothed version of the magnitude spectrum of the input speech signal.

***The mel-frequency cepstral coefficients (MFCC)*** are based on the perceptually motivated mel-scale filter-bank. This representation is frequently used in speech recognition (Davis and Mermelstein [57], Rabiner and Juang [216]). The MFCC have been utilized for BWE in Enbom and Kleijn [64] and Nilsson and Kleijn [187].

The calculation of the MFCC vector for a signal frame is performed in several steps: first, a pre-emphasis filter is applied to the input speech. Then, the Fourier coefficients are calculated via windowing (Hamming window), zero-padding and DFT. The Fourier coefficients are combined into 31 filter-bank outputs according to the mel-scale filter-bank as, for example, defined by Davis and Mermelstein [57]. The inverse discrete cosine transform is applied to the logarithms of the 31 filter-bank outputs, which yields 31 cepstral coefficients. The first cepstral coefficients are finally combined in the MFCC feature vector. Note that the MFCC representation includes information on the gain or power of the input speech in the feature vector because the signal is not normalized during the calculation of the MFCC.

***The normed frame energy*** Energy-based criteria provide a quite robust indication for voice activity, for example, Rabiner and Schafer [217], at least if the signal-to-noise ratio is sufficiently high. The signal power further differs for distinct speech sounds: in general the short-term power of the signal is greater for voiced sounds, while it is lower for unvoiced sounds. To become independent of long-term variations of the signal power, the frame energy has to be normalized adaptively.

For the $m$th frame, the normed logarithmic frame energy may, for example, be calculated by

$$x_{\text{nrp}}(m) = \frac{\log E(m) - \log E_{\min}(m)}{\log \overline{E}(m) - \log E_{\min}(m)}, \tag{6.36}$$

with

$$E(m) = \sum_{\kappa=0}^{N_\kappa - 1} s_{\text{nb}}^2(\kappa)$$

$$E_{\min}(m) = \min_{\mu=0}^{N_{\min}} E(m-\mu)$$

$$\overline{E}(m) = \alpha\, \overline{E}(m-1) + (1-\alpha)\, E(m).$$

A reasonable setting for telephone speech sampled with $f_{\text{s}} = 8\,\text{kHz}$ and with a frame rate of 50 frames/sec is to use a forgetting factor of $\alpha = 0.96$ and a size of $N_{\min} = 200$ for the minimum search window.

***The gradient index*** has been proposed for the voiced/unvoiced classification of speech segments (Paulus [207, 206]). The measure is based on the sum of magnitudes of the gradient of the speech signal at each change of direction

$$x_{\text{gi}} = \sum_{\kappa=2}^{N_\kappa} \frac{\Psi(\kappa)\, |s_{\text{nb}}(\kappa) - s_{\text{nb}}(\kappa-1)|}{\sqrt{\frac{1}{N_\kappa} E(m)}}. \tag{6.37}$$

$\Psi(\kappa)$ is an indicator function for the 'change of direction' of the signal. That is, $\Psi(\kappa) = 1/2\,|\psi(\kappa) - \psi(\kappa-1)|$, where the variable $\psi(\kappa)$ denotes the sign of the gradient $s_{\text{nb}}(\kappa) - s_{\text{nb}}(\kappa-1)$, that is, $\psi(\kappa) \in \{-1, 1\}$.

***The zero-crossing rate*** counts the number of times the signal crosses the zero level within each frame (Sec. I.3.2.4 and Eqn. I.1). The zero-crossing rate has been used extensively in speech recognition (Rabiner and Schafer [217]) and as a voicing criterion (Atal and Rabiner [22], Campbell and Tremain [43], and Wang [294]).

***The pitch period*** in the current speech frame depends on the instantaneous fundamental frequency $F_0$ of the speech signal. Its calculation is based on the auto-correlation function of the signal (see above): the position of the local maximum of the ACF in a limited range of reasonable time lags, for example, corresponding to pitch frequencies between 60 and 400 Hz, gives the estimated pitch period (e.g. Hess [109], Vary *et al.* [286]).

***The kurtosis*** is a measure from higher-order statistics that is based on the fourth and second-order moments of the signal. Here, we use an estimate of the *local* kurtosis

(Krishnamachari [149])

$$x_{\mathrm{k}} = \frac{\frac{1}{N_\kappa}\sum_{\kappa=0}^{N_\kappa-1}\left(s_{\mathrm{nb}}(\kappa)\right)^4}{\left(\frac{1}{N_\kappa}\sum_{\kappa=0}^{N_\kappa-1}\left(s_{\mathrm{nb}}(\kappa)\right)^2\right)^2}. \tag{6.38}$$

The kurtosis is a measure of the 'Gaussianity' of a random signal, and it is a dimensionless parameter. A Gaussian random variable has a kurtosis of 3. In the short-term, it can be observed that the local kurtosis is less than 3 for most voiced speech sounds. There are substantial peaks in the local kurtosis measure at the onset of plosives and of strong vowels.

***The spectral centroid*** is defined as the 'centre of gravity' of the magnitude spectrum of the bandlimited speech, reflecting its 'brightness' (see Sec. 1.4.6 on timbre)

$$x_{\mathrm{sc}} = \frac{\sum_{i=0}^{N_i/2} i \cdot \left|S_{\mathrm{nb}}(\mathrm{e}^{j\Omega_i})\right|}{\left(\frac{N_i}{2}+1\right)\sum_{i=0}^{N_i/2}\left|S_{\mathrm{nb}}(\mathrm{e}^{j\Omega_i})\right|}. \tag{6.39}$$

The quantity $S_{\mathrm{nb}}(\mathrm{e}^{j\Omega_i})$ labels the $i$th coefficient of a *discrete Fourier transformation* (DFT) of the length $N_i$ of the input signal frame. The spectral centroid is mainly around 1500 Hz (corresponding to $x_{\mathrm{sc}} \approx 0.35$ for a sampling rate of 8 kHz) for voiced speech sounds and increases significantly for unvoiced speech segments (Heide and Kang [107], Abdelatty Ali *et al.* [17], Abdelatty Ali and Van der Spiegel [16]). Note that this definition of spectral centroid $x_{\mathrm{sc}}$ differs from the spectral centroid $C_S$ (Eqn. 1.95).

***The spectral flatness*** is defined as the ratio between the geometric and arithmetic mean of the estimated power spectrum

$$x_{\mathrm{sfm}} = \frac{\sqrt[N_i]{\prod_{i=0}^{N_i-1}\left|S_{\mathrm{nb}}(\mathrm{e}^{j\Omega_i})\right|^2}}{\frac{1}{N_i}\sum_{i=0}^{N_i-1}\left|S_{\mathrm{nb}}(\mathrm{e}^{j\Omega_i})\right|^2}. \tag{6.40}$$

Because the arithmetic mean of a sequence of non-negative values is always greater than (or equal to) its geometric mean, the spectral flatness is between zero and one. The spectral flatness has, for example, been used to measure the tonality of signal segments (Johnston [136]).

### 6.5.5 EVALUATION

In this section, feature extraction for the typical application of bandwidth extension of telephone speech will be considered. That is, the narrowband speech signal $s_{\mathrm{nb}}(k)$ has frequency components in the range of 0.3–3.4 kHz. By the BWE algorithm, a wideband signal $s_{\mathrm{wb}}(k)$ with frequency components from 50 Hz up to 7 kHz shall be produced. We distinguish between the applications of low-frequency extension (below 300 Hz) and high-frequency extension (3.4–7 kHz).

For measuring mutual information and separability, speech signals are sub-divided into frames with a length of 20 ms. For each signal frame, all of the primary features defined in Sec. 6.5.4 and (from corresponding wideband speech) the vector $\mathbf{y}$ is determined. The vector $\mathbf{y}$ consists of weighted cepstral coefficients representing the gain and shape of the spectral envelope within the missing frequency band (50–300 Hz respectively 3.4–7 kHz) according to Eqn. 6.22. All of the measurements were performed using the BAS SI100 speech corpus (Schiel [234]).

***Mutual information and separability*** The estimated mutual information and separabilities between $\mathbf{y}$ and the investigated primary features are listed in Table 6.1. It can be observed that the features describing the spectral envelope of the bandlimited speech in fact play a major role for the bandwidth extension. Both the mutual information and the separability measures are maximal for these features. It must be taken into account, however, that the dimension of the primary features from this group is ten times higher than those of the scalar features. The mutual information and separabilities for the MFCC

**Table 6.1** Estimates of mutual information $I(\mathbf{x};\mathbf{y})$ and separability $\zeta(\mathbf{x})$ for bandwidth extension of telephone speech (0.3–3.4 kHz). For calculating the separability, the 16 classes were defined by vector quantizing the true wideband spectral envelope representation $\mathbf{y}$ (Jax and Vary [133])

| Feature vector $\mathbf{x}$ | $\dim \mathbf{x}$ | Towards high frequencies | | Towards low frequencies | |
|---|---|---|---|---|---|
| | | $I(\mathbf{x};\mathbf{y})$ [bit/frame] | $\zeta(\mathbf{x})$ (16 classes) | $I(\mathbf{x};\mathbf{y})$ [bit/frame] | $\zeta(\mathbf{x})$ (16 classes) |
| ACF | 10 | 2.6089 | 1.6349 | 2.7530 | 2.3977 |
| LPC | 10 | 2.3054 | 1.5295 | 2.1100 | 1.7901 |
| LSF | 10 | 2.3597 | 1.5596 | 2.2125 | 2.5817 |
| LPC-cepstrum | 10 | 2.2401 | 1.4282 | 2.1778 | 2.3879 |
| Cepstrum | 10 | 2.3075 | 1.5483 | 1.9398 | 2.5473 |
| MFCC | 10 | 2.3325 | 2.2659 | 3.0771 | 6.6142 |
| ACF ( 1 ) | 1 | 0.7514 | 1.1237 | 0.7324 | 1.1065 |
| ACF ( pitch period ) | 1 | 0.4450 | 0.4058 | 0.5441 | 0.6745 |
| Frame energy | 1 | 0.9285 | 1.0756 | 1.3968 | 4.2328 |
| Gradient index | 1 | 0.8011 | 1.2520 | 0.5403 | 0.6983 |
| Zero-crossing rate | 1 | 0.7453 | 1.0795 | 0.7456 | 1.1685 |
| Pitch period | 1 | 0.2451 | 0.0530 | 0.4823 | 0.1122 |
| Local kurtosis | 1 | 0.2037 | 0.0225 | 0.2979 | 0.0809 |
| Spectral centroid | 1 | 0.7913 | 1.0179 | 0.6630 | 0.9276 |
| Spectral flatness | 1 | 0.4387 | 0.3538 | 0.4201 | 0.4648 |

feature vector are highest because this vector already incorporates some information on the power level of the narrowband speech.

To achieve the best results with the BWE algorithm, it can further be motivated from Table 6.1 to additionally include certain scalar features in the feature vector. Particularly, the consideration of the frame energy as well as the gradient index, zero-crossing rate, and/or spectral centroid is very promising.

***Linear discriminant analysis*** To evaluate the impact of a linear discriminant analysis, the estimation quality obtained with the transformed feature vectors was determined. The HMM-based MMSE estimation rule from Sec. 6.9 (Jax and Vary [133]) was used with $N_S = 64$ HMM states and $L = 16$ mixture components in the state-specific GMMs. Both speaker-dependent and speaker-independent models were investigated. The results are expressed in terms of the *root mean square log spectral distortion* (RMS LSD) of the estimated spectral envelope within the missing frequency band, according to the definition of the performance measure from Sec. 6.4.2. The 15-dimensional feature super-vector $\mathbf{x}_0$ consisted of the first 10 normalized auto-correlation coefficients, the zero-crossing rate, the normed frame energy, the gradient index, the local kurtosis, and the spectral centroid.

In Fig. 6.20, the mean performances that were obtained both without LDA and with the application of LDAs for the dimensions $b = 1 \ldots 5$ are depicted. As expected, the distortions of the estimates are decreased by increasing the dimension of the LDA transform. Remarkably, the achieved performances with a dimension of the LDA transform of $b = 5$ are even superior to those of the estimator that uses the original non-transformed feature vectors with a dimension of $\beta = 15$. This effect is the result of the improved compactness of the feature vectors: if the dimension of the feature vectors $\mathbf{x}$ is reduced significantly, the quality of the statistical modelling is enhanced. Thus, by utilizing a linear discriminant analysis, the performance and robustness of the bandwidth extension system can be improved, yet simultaneously reducing the computational complexity of the estimation algorithm substantially.

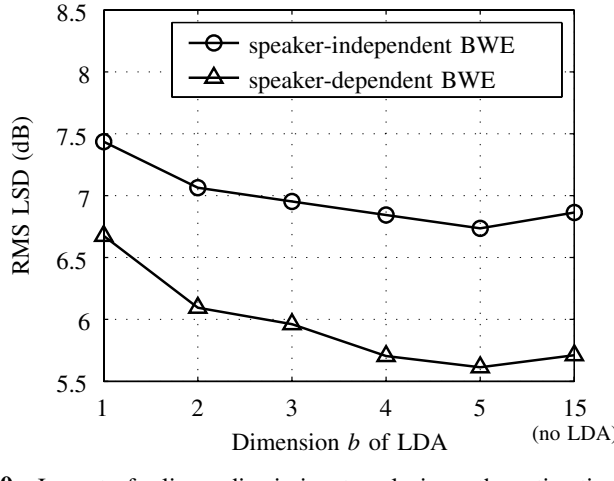



**Figure 6.20** Impact of a linear discriminant analysis on the estimation performance

## 6.6 CODEBOOK MAPPING

The first and most commonly used method for estimating the wideband spectral envelope is the codebook mapping approach (Carl and Heute [45], Yoshida and Abe [301], Carl [44], Yasukawa [299]). The principle of this class of algorithms is based on the observation that there occur only a limited number of typical sounds (i.e. typical shapes of the spectral envelope) in speech signals. Accordingly, the codebook mapping approach is based on a pair of coupled codebooks that contain representations of the spectral envelopes of the narrowband and wideband speech, respectively.

The basic algorithm is depicted in Fig. 6.21. For each signal frame, the spectral envelope of the narrowband speech signal, represented by the feature vector $\mathbf{x}$, is compared to a list of typical narrowband spectral envelopes that are stored in a pre-trained codebook. The most similar codebook entry is selected. In parallel to the searched primary narrowband codebook, there exists a second codebook, the so-called shadow codebook, which contains corresponding wideband spectral envelope representatives. Hence, the estimate $\tilde{\mathbf{y}}$ of the wideband spectral envelope is simply the entry of the shadow codebook that is assigned to the previously selected codebook entry of the narrowband codebook.

The estimates $\tilde{\mathbf{y}}$ are confined to the discrete entries $\hat{\mathbf{y}}$ of the shadow codebook. This is beneficial, on one hand, because it is guaranteed that the estimate yields stable LP synthesis filters in the BWE algorithm framework. On the other hand, the performance of the codebook mapping method is restricted by the number and quality of the entries in the corresponding codebooks.

To improve the performance of the codebook mapping approach, some authors have proposed to use interpolation methods. Then, instead of a simple table lookup, the estimate $\tilde{\mathbf{y}}$ is determined by a weighted sum of all or the most probable codebook entries

$$\tilde{\mathbf{y}} = \sum_{i=1}^{N_S} w_i \, \hat{\mathbf{y}}_i, \tag{6.41}$$

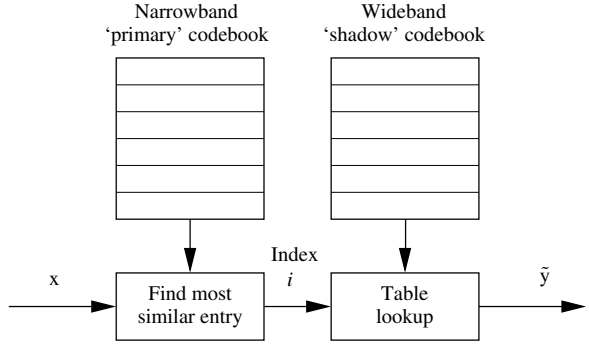



**Figure 6.21** Estimation of the spectral envelope representation $\mathbf{y}$ via codebook mapping. Corresponding entries with the same index in both codebooks reflect properties of the same typical speech sound

where $\hat{\mathbf{y}}_i$ denotes the $i$th entry of the shadow codebook. The weights $0 \leq w_i \leq 1$ have to be normed such that $\sum_{i=1}^{N_S} w_i = 1$. The individual weights $w_i$ are, for example, inverse proportional to the distance of the feature vector $\mathbf{x}$ to the respective codebook entry $\hat{\mathbf{x}}_i$ (Epps [65]).

### *6.6.1 VECTOR QUANTIZATION AND TRAINING OF THE PRIMARY CODEBOOK*

In the following paragraph, the entries of the primary narrowband codebook shall be defined by vector quantization (VQ) of the feature vector $\mathbf{x}$. Each code vector of the VQ represents the properties of a typical speech sound. For a comprehensive introduction into vector quantization see, for example, Gersho and Gray [87] and Gray and Neuhoff [97].

Vector quantization of the $b$-dimensional feature vectors $\mathbf{x} = [x_0, \ldots x_{b-1}]^{\mathrm{T}}$ is described by the mapping $Q : \mathbb{R}^b \rightarrow \mathcal{C}_{\mathbf{x}}$ from the $b$-dimensional Euclidian space into the finite sub-space $\mathcal{C}_{\mathbf{x}}$. This sub-space is defined by a codebook $\mathcal{C}_{\mathbf{x}}$ in which all possible representatives $\hat{\mathbf{x}}_i$ are combined, that is, $\mathcal{C}_{\mathbf{x}} = \{\hat{\mathbf{x}}_1, \ldots \hat{\mathbf{x}}_{N_S}\}$. The number of representatives is denoted by $N_S$.

The quantization mapping $Q$ is defined such as to minimize some error criterion $d(\mathbf{x}, \hat{\mathbf{x}}_i)$ between the input vectors $\mathbf{x}$ and the entries $\hat{\mathbf{x}}_i$, $i = 1 \ldots N_S$ of the vector codebook

$$Q(\mathbf{x}) = \arg \min_{\hat{\mathbf{x}}_i \in \mathcal{C}_{\mathbf{x}}} \mathrm{d}(\mathbf{x}, \hat{\mathbf{x}}_i). \tag{6.42}$$

Note that quantization with this mapping in general requires an extensive codebook search. There exist computationally more efficient, albeit sub-optimal, schemes that are commonly used for large codebooks, for example, *multi-stage vector quantization* (MSVQ) (Gersho [87]).

By the mapping from Eqn. 6.42, a region $\Upsilon_i$ of the $b$-dimensional Euclidian space, the quantizer cell, is assigned to each code vector $\hat{\mathbf{x}}_i$

$$\Upsilon_i = \left\{\mathbf{x} \in \mathbb{R}^b : \ \mathrm{d}(\mathbf{x}, \hat{\mathbf{x}}_i) < \ \mathrm{d}(\mathbf{x}, \hat{\mathbf{x}}_j), \forall \hat{\mathbf{x}}_j \in \mathcal{C}_{\mathbf{x}} \setminus \{\hat{\mathbf{x}}_i\}\right\}. \tag{6.43}$$

The set union of all quantizer regions $\Upsilon_i$ fills the entire Euclidian space $\mathbb{R}^b$ without overlappings, that is, $\bigcup_{i=1}^{N_S} \Upsilon_i = \mathbb{R}^b$, and $\Upsilon_i \cap \Upsilon_j = \emptyset$ for any $j \neq i$.

During training of the vector quantizer, the objective is to minimize the mean quantization distortion. For a fixed number $N_S$ of code vectors, this is accomplished by modifying the code vectors according to

$$\begin{aligned} \mathcal{C}_{\mathbf{x}} &= \arg \min_{\mathcal{C}} E\left\{ \mathrm{d}(\mathbf{x}, Q(\mathbf{x}))\right\} \\ &\approx \arg \min_{\mathcal{C}} \frac{1}{N_m} \sum_{m=0}^{N_m - 1} \min_{\hat{\mathbf{x}} \in \mathcal{C}} \mathrm{d}(\mathbf{x}(m), \hat{\mathbf{x}}). \end{aligned} \tag{6.44}$$

This task is usually performed by an iterative refinement procedure based on a large set of training vectors $\mathbf{x}(m)$, $m = 0 \ldots N_m - 1$. Commonly, the well-known LBG algorithm (Linde *et al.* [162]) is used, which is a variant of the generalized Lloyd algorithm (Lloyd [163]). By the training, a clustering of the training data is obtained.

### 6.6.2 TRAINING OF THE SHADOW CODEBOOK

Because there is a fixed relationship between the entries of the primary codebook and of the shadow codebook, the shadow codebook cannot be trained until having obtained the primary codebook. The entries $\hat{\mathbf{y}}_i$ of the shadow codebook $\mathcal{C}_\mathbf{y}$ are then defined by clustering of the training data w.r.t. the primary codebook $\mathcal{C}_\mathbf{x}$.

Since the mapping between $\mathbf{x}$ and $\mathbf{y}$ is not unique, the 'quantizer cells', defined by VQ of $\mathbf{x}$ but situated in the Euclidian space of $\mathbf{y}$, will in general be overlapping. Individual cells may even be discontiguous. Accordingly, to describe the regions assigned to the $i$th code vector $\hat{\mathbf{y}}_i$, it is not sufficient to use a minimum distortion criterion like in Eqn. 6.43. Instead, the $d$-dimensional conditional probability density function $p(\mathbf{y}|\mathbf{x} \in \Upsilon_i)$ with $\mathbf{y} = [y_0 \dots y_{d-1}]^\mathrm{T}$ has to be employed. With this, we can define the optimal code vectors as

$$\begin{aligned}\hat{\mathbf{y}}_i &= \arg\min_{\hat{\mathbf{y}}\in\mathbb{R}^d} E\left\{\, \mathrm{d}(\mathbf{y},\hat{\mathbf{y}}) \middle| \mathbf{x} \in \Upsilon_i \right\} \\ &= \arg\min_{\hat{\mathbf{y}}\in\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{y}|\mathbf{x} \in \Upsilon_i)\ \mathrm{d}(\mathbf{y},\hat{\mathbf{y}})\, \mathrm{d}\mathbf{y}. \end{aligned} \tag{6.45}$$

If, for example, the mean-square error criterion $d(\mathbf{y},\hat{\mathbf{y}}) = \|\mathbf{y}-\hat{\mathbf{y}}\|^2$ is used, then Eqn. 6.45 is solved by the conditional expectation $\hat{\mathbf{y}}_i = E\{\mathbf{y}|\mathbf{x} \in \Upsilon_i\}$. The $i$th expectation can be determined using a large number of pairs of training vectors $\{\mathbf{x}(m), \mathbf{y}(m)\}$, $m = 1 \dots N_m$ by averaging the vectors $\mathbf{y}$ extracted from those signal frames for which $\mathbf{x}(m) \in \Upsilon_i$.

The performance of the vector quantization approach strongly depends on the choice of representations of $\mathbf{x}$ and $\mathbf{y}$, and on the chosen distortion measures $d(\mathbf{x},\hat{\mathbf{x}})$ respectively $d(\mathbf{y},\hat{\mathbf{y}})$. Since in most implementations of the codebook mapping method for BWE the feature vector $\mathbf{x}$ reflects some representation of the spectral envelope of the narrowband speech, usually distortion measures used in speech coding for optimizing the quantization of LPC coefficients are employed, for example, Gray and Markel [95] and Gray *et al.* [96]. In principle, it is not necessary that the distortion measures for the primary and shadow codebook are identical.

Besides the distortion measure(s), the performance of codebook mapping depends on the sizes of the primary and/or shadow codebook. The estimation distortion is lower, the higher the number of codebook entries. Several authors have found that the codebook mapping performance in bandwidth extension for telephone speech saturates for codebook sizes greater than about 256 (Carl [44], Epps [65]). In Epps [65], algorithms have been developed that allow to decrease the size of the shadow codebook (with a fixed primary codebook) without sacrificing lots of performance.

## 6.7 LINEAR MAPPING

Another approach that has been used successfully to estimate the representation $\mathbf{y}$ of the wideband spectral envelope from a feature vector $\mathbf{x}$ is by linear mapping or piecewise-linear mapping, for example, Nakatoh *et al.* [183], Epps and Holmes [66], and Miet *et al.* [174]. With linear mapping, an estimate of the unknown quantity $\mathbf{y}$ is derived from the

observed feature vector $\mathbf{x}$ by the transformation

$$\tilde{\mathbf{y}} = \mathbf{A}^{\mathrm{T}} \cdot \mathbf{x}. \tag{6.46}$$

The dimension of the transformation matrix $\mathbf{A}$ is $b \times d$ with $b = \dim \mathbf{x}$ and $d = \dim \mathbf{y}$. The whole a priori knowledge on dependencies between $\mathbf{x}$ and $\mathbf{y}$ is contained in the matrix $\mathbf{A}$, which is derived and stored during the off-line training phase of the BWE system (see below). Therefore, there are no large memory requirements with this approach. Further, the estimation rule in Eqn. 6.46 is very simple to implement and computationally efficient.

A problem of the linear mapping algorithm is that it is in general not possible to strictly confine the estimates $\tilde{\mathbf{y}}$ to a reasonable and admissible range of values. Accordingly, mainly depending on the chosen representations of $\mathbf{x}$ and $\mathbf{y}$, applying the linear mapping rule sometimes results in an instability of the LP synthesis filter in the BWE system.

To prevent strong artefacts, such severe estimation errors are commonly concealed by mostly heuristically derived countermeasures. For example, in Chennoukh *et al.* [50], wideband LSF vectors ($\mathbf{y}$) are estimated from the narrowband LSF vectors ($\mathbf{x}$). This sometimes results in LSF elements larger than $\pi$, which makes it necessary to scale down the estimated vector so far that all elements are well below $\pi$ (cf. Sec. 6.4.1). Although preventing the worst, such measures impair the mean estimation performance.

Another flavour of linear mapping was used in Avendano *et al.* [23]. This contribution has the distinctive feature that the estimate $\tilde{\mathbf{y}}(m)$ depends not only on the features $\mathbf{x}(m)$ extracted from the current frame $m$ but also on the features from a number of preceding and following signal frames. In this case, the concept of linear mapping corresponds to multi-dimensional filtering.

### *6.7.1 TRAINING PROCEDURE*

During the training phase of the bandwidth extension system, the transformation matrix $\mathbf{A}$ has to be found. For this purpose, a database of true wideband speech is needed. By band-pass filtering, the corresponding narrowband speech is produced, and both signals are cut into time-aligned signal frames. From each pair of wideband and narrowband signal frames, the vectors $\mathbf{y}$ and $\mathbf{x}$ are extracted and collected in two large matrices $\mathbf{F}_{\mathbf{y}}$ and $\mathbf{F}_{\mathbf{x}}$. The rows of the matrix $\mathbf{F}_{\mathbf{y}}$ consist of the $d$-dimensional training vectors $\mathbf{y}(m)$, $m = 0 \ldots N_m$ computed from the wideband speech, and the rows of $\mathbf{F}_{\mathbf{x}}$ consist of $b$-dimensional feature vectors $\mathbf{x}(m)$ extracted from the corresponding narrowband speech frames

$$\mathbf{F}_{\mathbf{y}} = \begin{pmatrix} \mathbf{y}^{\mathrm{T}}(0) \\ \mathbf{y}^{\mathrm{T}}(1) \\ \vdots \\ \mathbf{y}^{\mathrm{T}}(N_m - 1) \end{pmatrix} \quad \text{and} \quad \mathbf{F}_{\mathbf{x}} = \begin{pmatrix} \mathbf{x}^{\mathrm{T}}(0) \\ \mathbf{x}^{\mathrm{T}}(1) \\ \vdots \\ \mathbf{x}^{\mathrm{T}}(N_m - 1) \end{pmatrix}. \tag{6.47}$$

The number $N_m$ denotes the number of signal frames in the training data set.

Now, the transformation matrix $\mathbf{A}$ shall be optimized such as to minimize the model error $\mathbf{y} - \mathbf{A}^{\mathrm{T}}\mathbf{x}$ for the complete training data set. This training procedure results from a

least squares approach, leading to minimization of the trace criterion

$$e^2 = \mathrm{tr}\left[(\mathbf{F_y} - \mathbf{F_x}\mathbf{A})^{\mathrm{T}}(\mathbf{F_y} - \mathbf{F_x}\mathbf{A})\right]. \tag{6.48}$$

This is the Frobenius norm of the error $\mathbf{F_y} - \mathbf{F_x}\mathbf{A}$, namely the sum of squares of all differences (estimation errors) $y_i(m) - \tilde{y}_i(m)$. Derivation with respect to one element $a_{ij}$ of the transformation matrix $\mathbf{A}$ delivers (Scharf [233])

$$\frac{\partial e^2}{\partial a_{ij}} = -2\,\mathrm{tr}\left[(\mathbf{F_y} - \mathbf{F_x}\,\mathbf{A})^{\mathrm{T}}\mathbf{F_x}\frac{\partial \mathbf{A}}{\partial a_{ij}}\right]. \tag{6.49}$$

For the minimization of $e^2$, the zero point of the derivative (Eqn. 6.49) has to be found, that is, Eqn. 6.49 has to be solved for the unknown matrix $\mathbf{A}$. This leads to the condition

$$(\mathbf{F_y} - \mathbf{F_x}\mathbf{A})^{\mathrm{T}}\mathbf{F_x} \equiv 0. \tag{6.50}$$

Because this condition is independent of the position of the element $a_{ij}$, solving it is sufficient for finding all elements of the transformation matrix $\mathbf{A}$.

A unique solution for the least squares problem can be found if and only if the inverse of the Gram matrix $\mathbf{F_x^T}\,\mathbf{F_x}$ exists[6]. Then, we arrive at the training algorithm

$$\mathbf{A} = \left(\mathbf{F_x^T}\,\mathbf{F_x}\right)^{-1}\mathbf{F_x^T}\,\mathbf{F_y}. \tag{6.51}$$

Note that the calculation of Eqn. 6.51 is not trivial. It requires computation and inversion of the Gram matrix $\mathbf{F_x^T}\,\mathbf{F_x}$, which is a huge $N_m \times N_m$ square matrix.

To verify the solution of Eqn. 6.51, we have to investigate the second derivative of the trace criterion (Eqn. 6.48). With $\frac{\partial^2 \mathbf{A}}{\partial a_{ij}^2} = 0$, we get (Scharf [233])

$$\frac{\partial}{\partial a_{ij}}\left(\frac{\partial e^2}{\partial a_{ij}}\right) = \mathrm{tr}\left[\left(\mathbf{F_x}\frac{\partial \mathbf{A}}{\partial a_{ij}}\right)^{\mathrm{T}}\mathbf{F_x}\frac{\partial \mathbf{A}}{\partial a_{ij}}\right] \geq 0 \tag{6.52}$$

which is the Frobenius norm of the matrix $\mathbf{F_x}\frac{\partial \mathbf{A}}{\partial a_{ij}}$ and therefore always non-negative. Accordingly, we can be sure that the solution of Eqn. 6.51 indeed minimizes Eqn. 6.48.

### *6.7.2 PIECEWISE-LINEAR MAPPING*

The basic problem of the linear mapping approach described above is that the statistical model of a linear dependency between feature vector $\mathbf{x}$ and spectral envelope representation $\mathbf{y}$ in general is too simple to describe the true relationship. Consequently, the linear mapping approach has been extended by a preceding classification stage (Nakatoh *et al.* [183], Chennoukh *et al.* [50]), to better reflect the possibly non-linear relationship between $\mathbf{x}$ and $\mathbf{y}$: first, a classification of the feature vector $\mathbf{x}(m)$ is performed. Then, if the $i$th

[6] The inverse of the Gram matrix $\mathbf{F_x^T}\,\mathbf{F_x}$ exists iff the columns of $\mathbf{F_x}$ are mutually independent.

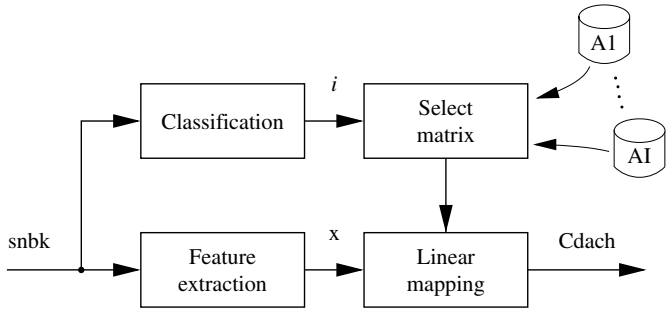



**Figure 6.22** Estimation of the wideband spectral envelope representation $\mathbf{y}$ via piecewise-linear mapping

class has been detected, a specific matrix $\mathbf{A}_i$ is used to determine the estimated spectral envelope representation $\tilde{\mathbf{y}}$ by linear mapping as described in Eqn. 6.46. The method is illustrated in Fig. 6.22.

For the first stage in Fig. 6.22, any classification method can be used. In literature, for example, vector quantization of the feature vector $\mathbf{x}$ has been utilized in Nakatoh *et al.* [183], or it was aimed at detecting certain phonemes via thresholding the reflection factors of the narrowband speech in Chennouk *et al.* [50]. Instead of a hard classification, the first stage in Fig. 6.22 can be enhanced to a soft decision scheme. Then, the final estimate $\tilde{\mathbf{y}}$ is determined by a weighted sum of individual mappings obtained for all classes (Nakatoh *et al.* [183]).

For the training of a piecewise-linear mapping method, the mapping matrix $\mathbf{A}_i$ for the $i$th class is computed by Eqn. 6.51 using only those signal frames from the training data set for which the $i$th class has been detected by the classification rule. It appears advantageous that the same classification rule is used during the training and application phase of the BWE system. Otherwise, there might be a model mismatch in the class-specific linear mapping matrices $\mathbf{A}_i$.

## 6.8 GAUSSIAN MIXTURE MODEL

The linear mapping approach described above has the disadvantage that the statistical model is principally limited to multivariate normal distributions. To employ more sophisticated statistically optimized estimation schemes, a more exact model of the joint PDF $p(\mathbf{x}, \mathbf{y})$, describing the joint behaviour of the two multi-dimensional random variables $\mathbf{x}$ and $\mathbf{y}$, is necessary (Park and Kim [200], Raza and Chan [220]). Then, even non-linear dependencies of $\mathbf{x}$ and $\mathbf{y}$ may be exploited.

Because the dimension $\dim \mathbf{x} + \dim \mathbf{y}$ of the joint PDF $p(\mathbf{x}, \mathbf{y})$ may be fairly large, prohibiting the use of histograms owing to memory constraints, it is necessary in practice to approximate the PDF by some parametric model $\tilde{p}(\mathbf{x}, \mathbf{y})$. Then, only the model parameters have to be obtained and stored in the off-line training phase.

For the definition of the joint PDF in the following, the two vectors $\mathbf{x} = [x_0, \ldots x_{b-1}]^{\mathrm{T}}$ and $\mathbf{y} = [y_0, \ldots y_{d-1}]^{\mathrm{T}}$ shall be combined in the column vector $\mathbf{z} = [\mathbf{x}^{\mathrm{T}}\, \mathbf{y}^{\mathrm{T}}]^{\mathrm{T}}$. A common

way to model unknown high-dimensional real-world probability density functions is the approximation with Gaussian mixture models (GMM, see e.g. Reynolds and Rose [222], Vaseghi [287]). In these parametric models, the PDF is approximated by the sum of weighted multivariate Gaussian distributions

$$p(\mathbf{x}, \mathbf{y}) \approx \tilde{p}(\mathbf{x}, \mathbf{y}) = \tilde{p}(\mathbf{z}) = \sum_{l=1}^{L} \rho_l \, \mathcal{N}(\mathbf{z}; \mu_{\mathbf{z},l}, \mathbf{V}_{\mathbf{z},l}) \tag{6.53}$$

with mean vectors $\mu_{\mathbf{z},l}$ and covariance matrices $\mathbf{V}_{\mathbf{z},l}$. The individual $(b+d)$-dimensional joint Gaussian densities (with $\dim \mathbf{z} = b + d$, $b = \dim \mathbf{x}$, $d = \dim \mathbf{y}$) are given by

$$\mathcal{N}(\mathbf{z}; \mu_{\mathbf{z},l}, \mathbf{V}_{\mathbf{z},l}) = \frac{\sqrt{\det \mathbf{A}_{\mathbf{z},l}}}{(2\pi)^{(b+d)/2}} \exp\left(-\frac{1}{2}(\mathbf{z} - \mu_{\mathbf{z},l})^{\mathrm{T}} \mathbf{A}_{\mathbf{z},l} \, (\mathbf{z} - \mu_{\mathbf{z},l})\right), \tag{6.54}$$

where $\mu_{\mathbf{z},l} = [\mu_{\mathbf{x},l}^{\mathrm{T}} \mu_{\mathbf{y},l}^{\mathrm{T}}]^{\mathrm{T}}$ and $\mathbf{A}_{\mathbf{z},l}$ is the inverse of the covariance matrix $\mathbf{V}_{\mathbf{z},l}$

$$\mathbf{A}_{\mathbf{z},l} = \begin{pmatrix} \mathbf{A}_{\mathbf{xx},l} & \mathbf{A}_{\mathbf{xy},l} \\ \mathbf{A}_{\mathbf{yx},l} & \mathbf{A}_{\mathbf{yy},l} \end{pmatrix} = \mathbf{V}_{\mathbf{z},l}^{-1} = \begin{pmatrix} \mathbf{V}_{\mathbf{xx},l} & \mathbf{V}_{\mathbf{xy},l} \\ \mathbf{V}_{\mathbf{yx},l} & \mathbf{V}_{\mathbf{yy},l} \end{pmatrix}^{-1}. \tag{6.55}$$

The scalar weighting factors $\rho_l$ in Eqn. 6.53 define the relative contribution of the $l$th Gaussian distribution to the modelled PDF. The model represents a true PDF if the weighting factors meet the constraints

$$0 \leq \rho_l \leq 1 \qquad \text{and} \qquad \sum_{l=1}^{L} \rho_l = 1. \tag{6.56}$$

The parameters of the GMM are combined in the set $\Theta = \{\Theta_l;\ l = 1, 2, \ldots L\}$ with the subsets $\Theta_l = \{\rho_l, \mu_{\mathbf{z},l}, \mathbf{V}_{\mathbf{z},l}\}$ of the respective parameters of the individual Gaussian mixture components.

It has been shown that any smooth continuous probability density function can be approximated arbitrarily closely by increasing the model order $L$ (Sorenson [254]). To show the qualitative behaviour, Gaussian mixture models with different orders $L$ for an exemplary two-dimensional random process are illustrated in Fig. 6.23.

### *6.8.1 MINIMUM MEAN SQUARE ERROR ESTIMATION*

The *minimum mean square error* (MMSE) estimation rule shall take the joint PDF $p(\mathbf{x}, \mathbf{y})$ into consideration. The aim of the MMSE criterion is the minimization of

$$\begin{aligned} \mathcal{D}_{\text{MSE}}(\mathbf{y}, \tilde{\mathbf{y}}|\mathbf{x}) &= E\left\{\|\mathbf{y} - \tilde{\mathbf{y}}\|^2 | \mathbf{x}\right\} \\ &= \int_{\mathbb{R}^d} p(\mathbf{y}|\mathbf{x}) \, \|\mathbf{y} - \tilde{\mathbf{y}}\|^2 \, \mathrm{d}\mathbf{y} \end{aligned} \tag{6.57}$$

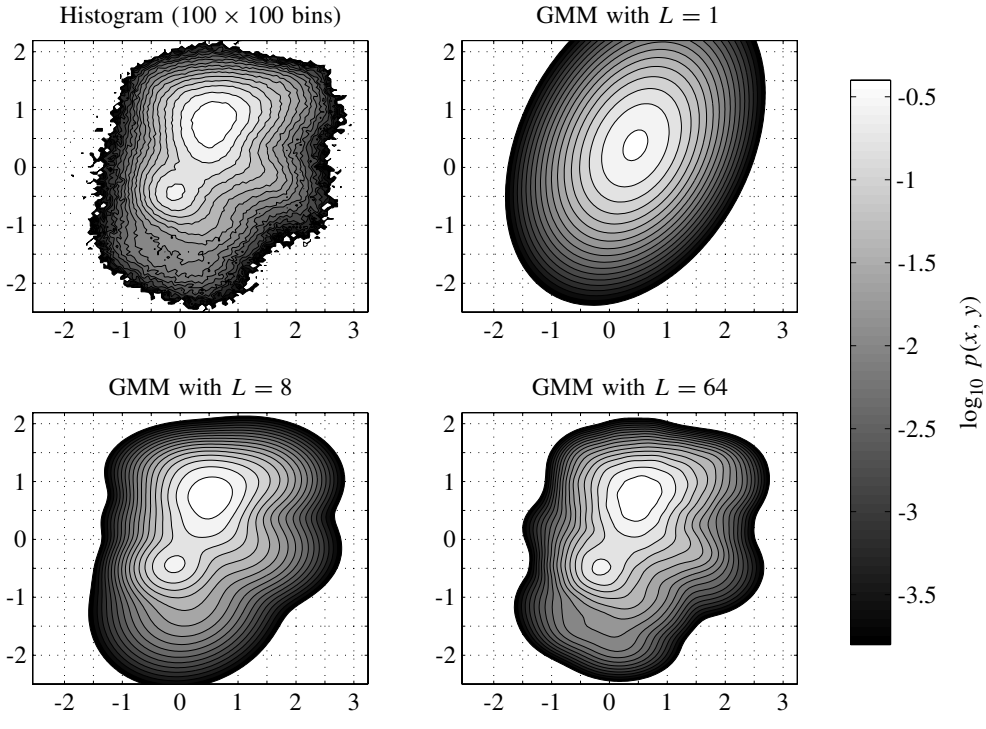



**Figure 6.23** Contour diagrams of Gaussian mixture models for two-dimensional exemplary data. In the upper left diagram, a measured histogram of the training data is shown. The other diagrams illustrate GMMs of this data with 1, 8, and 64 multivariate normal distributions, respectively

The solution is the conditional expectation

$$\tilde{\mathbf{y}} = E\{\mathbf{y}|\mathbf{x}\} = \int_{\mathbb{R}^d} \mathbf{y}\, p(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbf{y}, \tag{6.58}$$

which can be calculated in closed form from the parameters of a GMM of the joint PDF $p(\mathbf{x}, \mathbf{y})$ (Park and Kim [200], Jax [128])

$$\tilde{\mathbf{y}} = \sum_{l=1}^{L} \rho_{\mathbf{y}|\mathbf{x},l} \left( \mu_{\mathbf{y},l} - \left( (\mathbf{x} - \mu_{\mathbf{x},l})^{\mathrm{T}}\, \mathbf{A}_{\mathbf{xy},l}\, \mathbf{A}_{\mathbf{yy},l}^{-1} \right)^{\mathrm{T}} \right). \tag{6.59}$$

The 'new' weighting factors $\rho_{\mathbf{y}|\mathbf{x},l}$ are defined by

$$\rho_{\mathbf{y}|\mathbf{x},l} = \frac{\rho_l\, \mathcal{N}(\mathbf{x};\, \mu_{\mathbf{x},l},\, \mathbf{V}_{\mathbf{xx},l})}{\sum_{l=0}^{L} \rho_l\, \mathcal{N}(\mathbf{x};\, \mu_{\mathbf{x},l},\, \mathbf{V}_{\mathbf{xx},l})}. \tag{6.60}$$

The parameters $\mu_{\mathbf{x},l}$ and $\mathbf{V}_{\mathbf{xx},l}$ of the marginal Gaussian densities $\mathcal{N}(\mathbf{x}; \mu_{\mathbf{x},l}, \mathbf{V}_{\mathbf{xx},l})$ can be determined from the parameters of the GMM using the definitions from Eqn. 6.55.

Bandwidth extension using raw Gaussian mixture models has been introduced in Park and Kim [200]. GMMs are also often used as parts of *hidden Markov models* (HMMs), which will be treated in the next section.

### *6.8.2 TRAINING BY THE EXPECTATION-MAXIMIZATION ALGORITHM*

The parameters of the Gaussian mixture model are determined and stored during an off-line training phase. For the training of a GMM, a variety of algorithms have been proposed, which are based on different optimization criteria, for example, Bahl *et al.* [24], Gopalakrishnan *et al.* [94], Valtchev *et al.* [280], Schlüter and Macherey [235], Hedelin and Skoglund [106], Povey and Woodland [212], and Yang and Zwolinksi [298]. Most of these training algorithms have been applied to the training of statistical models for speech recognition.

Here, the *expectation-maximization* (EM) algorithm shall be outlined, which is prevalent in the GMM literature (e.g. Dempster *et al.* [59], Reynolds and Rose [222], Moon [175], Vaseghi [287]). The optimization criterion in the EM algorithm is the maximization of the log-likelihood function

$$\mathcal{L}(\Theta) = \log\left(\prod_{\mathbf{z}\in\Xi} \tilde{p}(\mathbf{z};\Theta)\right) = \sum_{\mathbf{z}\in\Xi} \log\left(\sum_{l=1}^{L} \rho_l \, \mathcal{N}(\mathbf{z};\Theta_l)\right). \tag{6.61}$$

Consequently, the method realizes a maximum likelihood (ML) optimization of the parameters $\Theta$ of the model, corresponding to a minimization of the Kullback-Leibler distance between the PDF $p(\mathbf{x}, \mathbf{y})$ and its model $\tilde{p}(\mathbf{x}, \mathbf{y})$ (Cover and Thomas [52]). The training of the GMM is based on a set $\Xi$ of training vectors that are taken from the original random process $\{\mathbf{x}(m), \mathbf{y}(m);\ m = 1 \ldots N_m\}$. The number of data vectors in the training set is denoted by $N_\Xi = |\Xi|$.

Unfortunately, the log-likelihood term (Eqn. 6.61) contains the logarithm of a sum such that a closed-form analytical solution for the maximization of $\mathcal{L}(\Theta)$ cannot be formulated. Instead, the EM approach leads to an iterative numerical training algorithm. The parameters $\rho_l$, $\mu_{\mathbf{z},l}$, and $\mathbf{V}_{\mathbf{z},l}$ of the GMM are refined in each iteration step (with the iteration index $\nu$) by the following update equations

$$\rho_l^{(\nu+1)} = \frac{1}{N_\Xi} \sum_{\mathbf{z}\in\Xi} \psi_l^{(\nu)}(\mathbf{z}),$$

$$\mu_{\mathbf{z},l}^{(\nu+1)} = \frac{\sum_{\mathbf{z}\in\Xi} \psi_l^{(\nu)}(\mathbf{z}) \cdot \mathbf{z}}{\sum_{\mathbf{z}\in\Xi} \psi_l^{(\nu)}(\mathbf{z})},$$

$$\mathbf{V}_{\mathbf{z},l}^{(\nu+1)} = \frac{\sum_{\mathbf{z}\in\Xi} \psi_l^{(\nu)}(\mathbf{z}) \cdot \left(\mathbf{z} - \mu_{\mathbf{z},l}^{(\nu+1)}\right)\left(\mathbf{z} - \mu_{\mathbf{z},l}^{(\nu+1)}\right)^{\mathrm{T}}}{\sum_{\mathbf{z}\in\Xi} \psi_l^{(\nu)}(\mathbf{z})}. \tag{6.62}$$

The variable $\psi_1(\mathbf{z})$ is defined by the a posteriori probability

$$\psi_1^{(\nu)}(\mathbf{z}) = \frac{\rho_l^{(\nu)} \mathcal{N}(\mathbf{z}; \mu_{\mathbf{z},l}^{(\nu)}, \mathbf{V}_{\mathbf{z},l}^{(\nu)})}{\sum_{l=1}^{L} \rho_l^{(\nu)} \mathcal{N}(\mathbf{z}; \mu_{\mathbf{z},l}^{(\nu)}, \mathbf{V}_{\mathbf{z},l}^{(\nu)})}. \tag{6.63}$$

For a detailed derivation of these terms, refer to the literature, for example, Dempster *et al.* [59], Moon [175], and Vaseghi [287]. Note that by simple modification of the above update rules, certain structures of the model parameters can be enforced, for example, diagonal covariance matrices.

For the initialization of the model prior to applying the EM algorithm, the training data set is sub-divided into clusters, for example, using the well-known binary-split LBG algorithm (Linde *et al.* [162], cf. Sec. 6.6.1). The centroids and covariances of the feature vectors that are assigned to the individual clusters are then used as the initial parameters $\mu_{\mathbf{z},l}^{(0)}$ and $\mathbf{V}_{\mathbf{z},l}^{(0)}$ of the model. The weighting factors $\rho_l^{(0)}$ shall be proportional to the number of feature vectors in the $l$th cluster.

It is a property of the EM algorithm that, provided the same large training data set is used for each iteration, the log-likelihood function increases strictly monotonically with every iteration step of the EM algorithm (Dempster *et al.* [59], Vaseghi [287]), that is, $\mathcal{L}(\Theta^{(\nu+1)}) \geq \mathcal{L}(\Theta^{(\nu)})$. The training is continued until the relative increase of the log-likelihood between two iterations falls below a predefined value $\varepsilon$, that is, the stop condition is

$$\frac{\mathcal{L}(\Theta^{(\nu+1)}) - \mathcal{L}(\Theta^{(\nu)})}{\left|\mathcal{L}(\Theta^{(\nu)})\right|} \leq \varepsilon. \tag{6.64}$$

Owing to the monotonical increase of the log-likelihood function during the training, it is guaranteed that the EM algorithm approaches a local maximum of $\mathcal{L}(\Theta)$. However, it can, in general, not be guaranteed that the *global* maximum is found.

## 6.9 HIDDEN MARKOV MODEL

In this section, the source-filter model from Sec. 6.2 shall be extended by a *hidden Markov model* (HMM). A HMM is a discrete-time *composite source model* (CSM), consisting of a finite number of independent sub-sources that are controlled by a switch, compare Fig. 6.24 (a). Each setting of the switch defines a state of the model, and for each state of the model the statistical properties of the output signal of the CSM correspond to the statistical properties of the selected sub-source. It is further assumed that the position of the switch is governed by a Markov chain. In this case, the model is referred to as a hidden Markov model in literature. The statistical characteristics of the state sequence can be described by a matrix of transition probabilities. Further details on hidden Markov models can be found in the literature, for example, Rabiner [215], Papoulis [199], Rabiner and Juang [216], and Vaseghi [287].

The application of a hidden Markov model to the process of speech production is illustrated in Fig. 6.24 (a). The state $\mathcal{S}_i$ of the source is represented by a switch that redirects

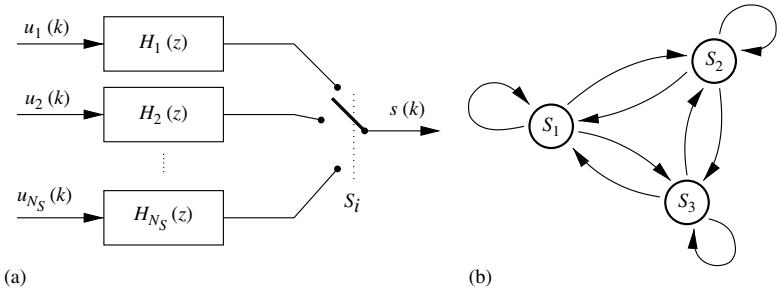



**Figure 6.24** (a) Hidden Markov model of the process of speech generation. The AR filters $H_i(z)$ and excitation signals $u_i(k)$ represent typical (wideband) speech sounds for each state. (b) State transition diagram of an ergodic first-order Markov chain with $N_S = 3$ states

the output of one of the sub-sources to the output $s(k)$ of the HMM. The sub-sources of the HMM correspond to individual source-filter models as introduced in Sec. 6.2.1. Each of the sub-sources is assumed to be *stationary* and represents the characteristics of one particular *wideband* speech sound. To simplify matters, state transitions are only allowed at the boundaries of signal frames in the following. Further, the transition from any state to any other state shall be possible, in which case the HMM is called ergodic[7].

For the bandwidth extension application, the states of the HMM are defined by vector quantization of the spectral envelope representation $\mathbf{y}$ (Jax *et al.* [129]). Every state $\mathcal{S}_i$ of the HMM corresponds to one entry $\hat{\mathbf{y}}_i$ of the VQ codebook $\mathcal{C}_\mathbf{y} = \{\hat{\mathbf{y}}_1 \ldots \hat{\mathbf{y}}_{N_S}\}$ such that the number of states $N_S$ in the HMM is the same as the number of entries in the codebook. The training of the VQ codebook is performed off-line with real wideband speech, using the LBG algorithm (Linde [162]).

Per definition, if wideband speech is available, the *true* state of the source in the $m$th frame can be determined by vector quantization of the spectral envelope representation $\mathbf{y}(m)$

$$\mathcal{S}_{\text{true}}(m) = \mathcal{S}_{i_{\text{opt}}(m)} \quad \text{where} \quad i_{\text{opt}}(m) = \arg\min_{i=1}^{N_S} \|\mathbf{y}(m) - \hat{\mathbf{y}}_i\|^2, \tag{6.65}$$

The true state sequence is needed during the off-line training phase to obtain the parameters of the statistical model describing the Markov states.

Note that, with the above definition, the Markov state is *not hidden* during training. Nevertheless, wideband speech is only available in the training phase whereas in the application phase of the BWE system the states $\mathcal{S}_i$ have to be identified from the features $\mathbf{x}$ of the narrowband speech. Corresponding estimators will be described in Sec. 6.9.2.

### *6.9.1 STATISTICAL MODEL OF THE MARKOV STATES*

For each possible state $\mathcal{S}_i$ of the hidden Markov model, the features $\mathbf{x}$ as well as the wideband spectral envelope representation $\mathbf{y}$ exhibit characteristic statistical properties.

[7] In certain applications, for example, for speech recognition, it is useful to restrict the possible state transitions to get directed state diagrams (e.g. a left–right model).

To describe these properties, a statistical model that consists of three parts is employed: the state and transition probabilities of the Markov chain, the probability density function (PDF) of the feature vectors $\mathbf{x}$ of the bandlimited speech (observation probability), and the PDF or expectation of the estimated quantity $\mathbf{y}$ (emission probability). We will elaborate on these three parts of the statistical model below. The information that is contained in the statistical model is needed as a priori knowledge for the subsequent estimation methods.

The training of the statistical model is performed off-line using a training database of corresponding wideband and bandlimited speech signals. The wideband speech is needed to calculate the true state sequence of the HMM. According to the true state $\mathcal{S}_{i_{\text{opt}}}$ of the HMM, as defined by Eqn. 6.65, the vectors $\mathbf{x}$ and $\mathbf{y}$ are assigned to the corresponding sets $\Xi_i$, $i = 1 \ldots N_{\mathcal{S}}$).

### Observation Probability

As the observation probability, we define the conditional probability density function $p(\mathbf{x}|\mathcal{S}_i)$ of the feature vectors $\mathbf{x}$. The conditioning is with respect to the state $\mathcal{S}$ such that there exists a separate PDF $p(\mathbf{x}|\mathcal{S}_i)$ for each HMM state $\mathcal{S}_i$. In accordance with the definition of the HMM, it is assumed that the observation $\mathbf{x}(m)$ for each frame only depends on the state $\mathcal{S}_{\text{true}}(m)$ of the Markov chain during that particular frame.

The modelling of the observation PDFs is complicated by the fact that the features $\mathbf{x}$ are continuous variables. Further, they constitute multi-dimensional vectors with the dimension $b$ for each signal frame. Therefore, we use parametric *Gaussian mixture models* (GMMs). The observation probability density is expressed by

$$p(\mathbf{x}|\mathcal{S}_i) \approx \tilde{p}(\mathbf{x}|\mathcal{S}_i) = \sum_{l=1}^{L} \rho_{il}\, \mathcal{N}(\mathbf{x};\, \mu_{il}, \mathbf{V}_{il}). \tag{6.66}$$

For each state $\mathcal{S}_i$ of the hidden Markov model, there exists an individual GMM with the parameter set $\Theta_i = \{\rho_{il}, \mu_{il}, \mathbf{V}_{il}; l = 1, 2, \ldots L\}$. The EM training procedure to determine the parameters $\Theta_i$ as well as some details on the structure and parameterization of GMMs has already been described in Sec. 6.8.

The observation probability constitutes the connection between the state of the HMM and the observed characteristics (features $\mathbf{x}$) of the bandlimited speech signal $s_{\text{nb}}(k)$. Consequently, the observation probability is the decisive element of the statistical model of the speech production process for detecting the momentary HMM state, and the modelling of $p(\mathbf{x}|\mathcal{S}_i)$ has to be implemented with special care.

### Emission Probability

The emission probability describes the statistical characteristics of the variable $\mathbf{y}$ representing the estimated spectral envelope of the missing frequency band. The modelling of the emission probability depends on the type of estimation rule that will be described in Sec. 6.9.2.

If the estimation of $\mathbf{y}$ is based solely on the detection of the actual state of the HMM, for example, by a ML (*maximum likelihood*), MAP (*maximum a posteriori*) or MMSE *soft* classification (MMSE variant I), the only information on $\mathbf{y}$ that can be extracted from

the conditional PDF $p(\mathbf{y}|\mathcal{S}_i)$ is the conditional expectation $E\{\mathbf{y}|\mathcal{S}_i\}$. Thus, it is sufficient to store the vectors $\hat{\mathbf{y}}_i = E\{\mathbf{y}|\mathcal{S}_i\}$ in a codebook. Note that this codebook is identical to the codebook $\mathcal{C}_\mathbf{y}$ of the constituting vector quantizer of the HMM.

In more sophisticated estimation rules, state-specific mutual dependencies between the variables $\mathbf{x}$ and $\mathbf{y}$ are additionally taken into account. For such estimators the emission probability is described by models of the conditional joint PDFs $p(\mathbf{x}, \mathbf{y}|\mathcal{S}_i)$. Since both $\mathbf{x}$ and $\mathbf{y}$ are multi-dimensional, continuous variables, state-specific Gaussian mixture models can be utilized to model these joint PDFs (compare Sec. 6.8).

### Parameters of the Markov Chain

The dependencies between the states of consecutive frames shall be considered in the statistical model. These dependencies are reproduced by the parameters of the Markov chain in the HMM. In the sequel, depending on the modelling and exploitation of the Markov chain parameters, the utilized a priori knowledge will be labelled as follows:

AK0: Only the probability of occurrence of the states is considered, that is, it is assumed that the probability of a state is independent from the state of the source at preceding or following frame instants.

AK1: A first-order ergodic Markov chain is assumed, that is, transition probabilities between consecutive states of the source are taken into account.

***State probability (AK0)*** The scalar value $P(\mathcal{S}_i)$ describes the (non-conditional) state probability, that is, the a priori probability that the HMM is in state $\mathcal{S}_i$ without incorporating any further observation or a priori knowledge, for example, of the feature vector $\mathbf{x}$, or of the state in preceding or following frames.

The state probabilities can easily be estimated by computing the true state sequence for the wideband training material and counting the number of occurrences $N_{\Xi_i} = |\Xi_i|$ of each state $\mathcal{S}_i$. The ratio between the number of occurrences of state $\mathcal{S}_i$ and the total number $N_m$ of speech frames in the training set gives the estimated state probability $\tilde{P}(\mathcal{S}_i) = N_{\Xi_i}/N_m$. The resulting probability values are stored in a table such that the actual bandwidth extension algorithm can later on access the a priori state probabilities by simple table lookups.

***Transition probabilities (AK1)*** The *transition probability* $P(\mathcal{S}_i(m+1)|\mathcal{S}_j(m))$ describes the conditional probability that the state of the source changes from state $\mathcal{S}_j$ in one signal frame to state $\mathcal{S}_i$ in the next frame.

Because the true state sequence is known during the training phase of the BWE algorithm, the transition probabilities can be estimated as the ratio between the counted number of occurrences of a particular transition from $\mathcal{S}_j$ to $\mathcal{S}_i$ and the total number of occurrences of state $\mathcal{S}_j$. Because, in general, transitions from any state to any other state are possible owing to the ergodicity of the Markov chain, a $N_\mathcal{S} \times N_\mathcal{S}$ matrix (i.e. a two-dimensional table) is necessary to store the transition probabilities for the later bandwidth extension.

Higher-order Markov modelling is possible, but the straightforward implementation yields huge lookup tables to store the transition probabilities, and the computational complexity of estimation rules increases exponentially with the model order.

### 6.9.2 ESTIMATION RULES

The actual classification or estimation constitutes the final step towards the determination of the coefficients $\tilde{\mathbf{y}}$ representing the spectral envelope of the missing frequency band. In this section, three well-known estimation methods are described, and their application in the context of the estimation of the spectral envelope is depicted. Two of the three methods are based on a posteriori probabilities that will first be defined in the next section. The most important advantage of the HMM-based estimation rules using a posteriori probabilities is that they explicitly use memory, that is they take observations from adjacent signal frames into account.

#### 6.9.2.1 Calculation of A Posteriori Probabilities

To be able to utilize a priori knowledge on the temporal dependencies of the states of the hidden Markov model, we define the *observation sequence* $\mathbf{X}(m_k) = \{\mathbf{x}(1) \ldots \mathbf{x}(m_k)\}$ containing all feature vectors that have been observed up to the $m_k$th frame. Note that the index $m_k$ of the most recently observed feature vector is allowed to be greater than the frame index $m$ of the currently processed signal frame, corresponding to an interpolation or look-ahead. The a posteriori probabilities shall be expressed with respect to all observed signal frames

$$P(\mathcal{S}_i(m)|\mathbf{X}(m_k)) = P(\mathcal{S}_i(m)|\mathbf{x}(1), \mathbf{x}(2), \ldots \mathbf{x}(m), \ldots \mathbf{x}(m_k)). \tag{6.67}$$

The definition and calculation of the a posteriori probabilities $P(\mathcal{S}_i(m)|\mathbf{X}(m_k))$ depends on the kind of a priori knowledge that shall be utilized. According to the definitions from Sec. 6.9.1, the two cases AK0 and AK1 will be distinguished in the following paragraphs.

***No consideration of transition probabilities (AK0)*** If only the state probabilities of the Markov chain shall be considered (AK0), it is assumed that the state of the source for the $m$th frame of the signal only depends on the features $\mathbf{x}(m)$ observed for that frame. Then, the a posteriori PMFs from Eqn. 6.67 can be simplified

$$P(\mathcal{S}_i(m)|\mathbf{x}(1), \mathbf{x}(2), \ldots \mathbf{x}(m), \ldots \mathbf{x}(m_k)) = P(\mathcal{S}_i(m)|\mathbf{x}(m)). \tag{6.68}$$

Applying Bayes' rule yields an expression for the a posteriori probabilities in which only the modelled observation probabilities and state probabilities are contained

$$\begin{aligned} P(\mathcal{S}_i(m)|\mathbf{x}(m)) &= \frac{p(\mathbf{x}(m)|\mathcal{S}_i(m))\, P(\mathcal{S}_i(m))}{p(\mathbf{x}(m))} \\ &= \frac{p(\mathbf{x}(m)|\mathcal{S}_i(m))\, P(\mathcal{S}_i(m))}{\sum_{j=1}^{N_\mathcal{S}} p(\mathbf{x}(m)|\mathcal{S}_j(m))\, P(\mathcal{S}_j(m))}. \end{aligned} \tag{6.69}$$

***First-order HMM (AK1)*** For a first-order hidden Markov model, the a posteriori PMF $P(\mathcal{S}_i(m)|\mathbf{X}(m_k))$ is expressed in terms of the joint PDF $p(\mathcal{S}_i(m), \mathbf{X}(m_k))$ and the PDF

$p(\mathbf{X}(m_k))$ of the observation sequence

$$P(\mathcal{S}_i(m)|\mathbf{X}(m_k)) = \frac{p(\mathcal{S}_i(m), \mathbf{X}(m_k))}{\sum_{j=1}^{N_\mathcal{S}} p(\mathcal{S}_j(m), \mathbf{X}(m_k))}. \tag{6.70}$$

The joint probability density function $p(\mathcal{S}_i(m), \mathbf{X}(m_k))$ can be determined as the product of the observation probability density of the current feature vector $\mathbf{x}(m)$ and of two components $\alpha_i(\cdot)$ and $\beta_i(\cdot)$ that comprise the contributions of the observed feature vectors from preceding and subsequent signal frames

$$p(\mathcal{S}_i(m), \mathbf{X}(m_k)) = \alpha_i(m)\, \beta_i(m)\, p(\mathbf{x}(m)|\mathcal{S}_i(m)). \tag{6.71}$$

The two quantities $\alpha_i(m)$ and $\beta_i(m)$ are calculated via forward and backward recursion, respectively, thereby utilizing the complete available observation sequence.

The quantity $\alpha_i(m)$ can be interpreted as the a priori probability $p(\mathcal{S}_i(m), \mathbf{X}(m-1))$ of the $i$th state of the HMM, considering all *past* observed feature vectors. The successive calculation of $\alpha_i(m)$ is based on the recursive equation

$$\alpha_i(m+1) = \sum_{j=1}^{N_\mathcal{S}} \alpha_j(m)\, p(\mathbf{x}(m)|\mathcal{S}_j(m))\, P(\mathcal{S}_i(m+1)|\mathcal{S}_j(m))$$

$$\alpha_i(1) = P(\mathcal{S}_i). \tag{6.72}$$

Because there exists no predecessor for the very first frame of the input speech, the recursion has to be initialized with the non-conditional state probabilities $P(\mathcal{S}_i)$. Owing to the recursive definition of $\alpha_i(m)$ in Eqn. 6.72, it is not necessary to store all past frames but it suffices to pass the a priori knowledge $\alpha_i(m+1)$ from one frame to the other.

If future observations shall also be taken into account for the a posteriori probability, that is, if $m_k > m$, the terms $\beta_i(m)$ can likewise be calculated recursively

$$\beta_i(m-1) = \sum_{j=1}^{N_\mathcal{S}} \beta_j(m)\, p(\mathbf{x}(m)|\mathcal{S}_j(m))\, P(\mathcal{S}_j(m)|\mathcal{S}_i(m-1)). \tag{6.73}$$

The initialization of the recursion has to be performed for the most recently observed signal frame with the index $m_k$ by $\beta_i(m_k) = 1$. If no future observations shall be considered by the estimation, the quantities $\beta_i(m)$ also have to be set to a value of $\beta_i(m) = 1$ in the calculation of the a posteriori probabilities (Eqn. 6.71).

#### 6.9.2.2 Maximum Likelihood Classification

A widely used classification method is the *maximum likelihood* (ML) approach. This estimation rule does not take the a priori knowledge on the state sequence into account.

That codebook entry of $\mathcal{C}_\mathbf{y}$ is selected which corresponds to that state of the HMM for which the observation density of the currently observed feature vector is maximized

$$\tilde{\mathbf{y}}_{\mathrm{ML}}(m) = E\left\{\mathbf{y}|\mathcal{S}_{i_{\mathrm{ML}}(m)}\right\} = \hat{\mathbf{y}}_{i_{\mathrm{ML}}(m)}$$

with

$$i_{\mathrm{ML}}(m) = \arg\max_{i=1}^{N_\mathcal{S}} p(\mathbf{x}(m)|\mathcal{S}_i(m)). \tag{6.74}$$

Consequently, the range of possible output values of the ML estimator is limited to the entries of the codebook. Note that the a posteriori probabilities from the previous section, that is, the state and transition probabilities of the HMM, are not utilized.

There are, in fact, certain parallels between the ML classification rule and the codebook mapping approach from Sec. 6.6: as with the primary codebook in Sec. 6.6, with Eqn. 6.74, fixed regions of the $b$-dimensional feature space are assigned to fixed pre-trained spectral envelope representatives[8]. The major difference is that the use of GMMs for $p(\mathbf{x}|\mathcal{S}_i)$ allows for a much more flexible definition of those regions, and the HMM-based training is more directly targeted on maximizing the estimation performance.

#### 6.9.2.3 Maximum A Posteriori Classification

The goal of the *maximum a posteriori* (MAP) rule is to maximize the a posteriori *probability mass function* (PMF). Accordingly, that entry of the codebook $\mathcal{C}_\mathbf{y}$ is selected which is assigned to the state of the HMM for which the a posteriori PMF $P(\mathcal{S}_i|\mathbf{X}(m_k))$ is maximum

$$\tilde{\mathbf{y}}_{\mathrm{MAP}}(m) = E\left\{\mathbf{y}|\mathcal{S}_{i_{\mathrm{MAP}}(m)}\right\} = \hat{\mathbf{y}}_{i_{\mathrm{MAP}}(m)}$$

with

$$i_{\mathrm{MAP}}(m) = \arg\max_{i=1}^{N_\mathcal{S}} P(\mathcal{S}_i(m)|\mathbf{X}(m_k)). \tag{6.75}$$

Because the normative factor in the denominator of the fraction in Eqn. 6.70 is identical for all states of the HMM, its value is irrelevant for the classification such that

$$i_{\mathrm{MAP}}(m) = \arg\max_{i=1}^{N_\mathcal{S}} p(\mathcal{S}_i(m), \mathbf{X}(m_k)). \tag{6.76}$$

In contrast to the ML approach, a priori knowledge about the state sequence of the HMM is utilized by the MAP method. The range of results of the estimation is, however, still limited to the contents of the codebook $\mathcal{C}_\mathbf{y}$. The MAP rule minimizes the number of mis-classifications of the HMM state.

[8] Actually, codebook mapping with Euclidian distance criterion (e.g. $d(\mathbf{x}, \hat{\mathbf{x}}) = \|\mathbf{x} - \hat{\mathbf{x}}\|^r$ with $r > 0$) can be interpreted as a special case of Eqn. 6.74 if the covariance matrices in $\tilde{p}(\mathbf{x}|\mathcal{S}_i)$ are fixed to $\mathbf{V}_{il} = \sigma^2\,\mathbf{I}$ in the EM algorithm.

#### 6.9.2.4 Minimum Mean Square Error Estimation

Now, the range of results of the estimation shall no longer be limited to the contents of the codebook $\mathcal{C}_{\mathbf{y}}$, but all values in the $d$-dimensional Euclidian space $\mathbb{R}^d$ are allowed. The aim of the *minimum mean square error* (MMSE) optimization is to minimize the error criterion

$$\begin{aligned}\mathcal{D}_{\text{MSE}}(\mathbf{y}(m), \tilde{\mathbf{y}}(m)|\mathbf{X}(m_k)) &= E\left\{\|\mathbf{y}(m) - \tilde{\mathbf{y}}(m)\|^2 \big| \mathbf{X}(m_k)\right\} \\ &= \int_{\mathbb{R}^d} p(\mathbf{y}(m)|\mathbf{X}(m_k))\, \|\mathbf{y}(m) - \tilde{\mathbf{y}}(m)\|^2 \, \mathrm{d}\mathbf{y}(m). \quad (6.77)\end{aligned}$$

The integral in Eqn. 6.77 has to be solved for the complete $d$-dimensional parameter space. The solution is the conditional expectation

$$\begin{aligned}\tilde{\mathbf{y}}_{\text{MMSE}}(m) &= E\left\{\mathbf{y}(m)\big|\mathbf{X}(m_k)\right\} \\ &= \int_{\mathbb{R}^d} \mathbf{y}(m)\, p(\mathbf{y}(m)|\mathbf{X}(m_k))\, \mathrm{d}\mathbf{y}(m). \quad (6.78)\end{aligned}$$

Because we do not have a model of the conditional PDF $p(\mathbf{y}(m)|\mathbf{X}(m_k))$ in closed form, this quantity has to be expressed indirectly via the states of the HMM

$$\begin{aligned}p(\mathbf{y}(m)|\mathbf{X}(m_k)) &= \sum_{i=1}^{N_{\mathcal{S}}} p(\mathbf{y}(m), \mathcal{S}_i(m)|\mathbf{X}(m_k)) \\ &= \sum_{i=1}^{N_{\mathcal{S}}} p(\mathbf{y}(m)|\mathcal{S}_i(m), \mathbf{x}(m))\, P(\mathcal{S}_i(m)|\mathbf{X}(m_k)). \quad (6.79)\end{aligned}$$

The second line in Eqn. 6.79 results from the model assumption that the vectors $\mathbf{x}(m)$ and $\mathbf{y}(m)$ exclusively depend on the state $\mathcal{S}(m)$ of the source in the $m$th signal frame. Inserting Eqn. 6.79 into Eqn. 6.78 yields the general state-based rule

$$\tilde{\mathbf{y}}_{\text{MMSE}}(m) = \sum_{i=1}^{N_{\mathcal{S}}} P(\mathcal{S}_i(m)|\mathbf{X}(m_k)) \int_{\mathbb{R}^d} \mathbf{y}(m)\, p(\mathbf{y}(m)|\mathcal{S}_i(m), \mathbf{x}(m))\, \mathrm{d}\mathbf{y}(m). \quad (6.80)$$

Depending on the available statistical model of the emission probability (compare Sec. 6.9), that is, whether only the state-specific expectation of $\mathbf{y}$ or the a priori knowledge on the joint PDF $p(\mathbf{x}, \mathbf{y}|\mathcal{S}_i)$ is at hand, two variants of MMSE estimators can be formulated. In addition, the AK0 and AK1 assumptions can be used. We will, however, not distinguish between these in the following.

***Variant I: 'soft classification'*** For the first variant of MMSE estimation, the emission probability shall be modelled without taking the observed feature vectors $\mathbf{x}$ into account.

By this, the conditioning on the feature vector $\mathbf{x}(m)$ within the integral on the right-hand side of Eqn. 6.80 is neglected, and the integral reflects the expectation of the coefficient vector $\mathbf{y}$ on the condition that the source is in the state $\mathcal{S}_i$

$$\int_{\mathbb{R}^d} \mathbf{y}(m)\; p(\mathbf{y}(m)|\mathcal{S}_i(m))\, \mathrm{d}\mathbf{y}(m) = E\{\mathbf{y}(m)|\mathcal{S}_i(m)\} = \hat{\mathbf{y}}_i. \tag{6.81}$$

The integral can be replaced by the corresponding entry $\hat{\mathbf{y}}_i$ of the pre-trained codebook. Substituting the a posteriori probability defined in Eqn. 6.67, we derive the MMSE classification rule

$$\tilde{\mathbf{y}}_{\text{MMSE}^{\text{I}}}(m) = \sum_{i=1}^{N_{\mathcal{S}}} \hat{\mathbf{y}}_i\; P(\mathcal{S}_i(m)|\mathbf{X}(m_k)). \tag{6.82}$$

Hence, the estimated coefficient set $\tilde{\mathbf{y}}_{\text{MMSE}^{\text{I}}}$ is calculated by the sum of the individual codebook entries that are weighted by the respective a posteriori probabilities of the corresponding states of the HMM. Accordingly, the described MMSE estimator can be interpreted as a *soft classification*. Note that with AK0 (with $\tilde{P}(\mathcal{S}_i) = 1/N_{\mathcal{S}}$) and fixed diagonal covariance matrices $\mathbf{V}_{il} = \sigma^2\,\mathbf{I}$, the MMSE rule (Eqn. 6.82) results in an MMSE-optimized codebook mapping approach with interpolation, compare Sec. 6.6.

***Variant II: 'cascaded estimation'*** The second variant of the MMSE estimation rule shall take the state-specific joint PDF $p(\mathbf{x}(m), \mathbf{y}(m)|\mathcal{S}_i(m))$ into consideration (Jax and Vary [132]). Then, the integral on the right-hand side of Eqn. 6.80 reflects the conditional expectation $E\{\mathbf{y}(m)|\mathcal{S}_i(m), \mathbf{x}(m)\}$. This conditional expectation can be calculated from the parameters of a Gaussian mixture model of the joint PDF $p(\mathbf{x}(m), \mathbf{y}(m)|\mathcal{S}_i(m))$ as described in Eqn. 6.59. Inserting the conditional expectation into Eqn. 6.80 leads to the second MMSE estimation rule

$$\tilde{\mathbf{y}}_{\text{MMSE}^{\text{II}}}(m) = \sum_{i=1}^{N_{\mathcal{S}}} E\{\mathbf{y}(m)|\mathcal{S}_i(m), \mathbf{x}(m)\}\; P(\mathcal{S}_i(m)|\mathbf{X}(m_k)). \tag{6.83}$$

This estimation rule can be interpreted as a cascaded estimation: first the state-dependent expectation of $\mathbf{y}$ is calculated for each state, followed by an individual weighting with the respective a posteriori probabilities.

Compared to the first variant of the MMSE estimation from Eqn. 6.82, the second variant (Eqn. 6.83) should always provide better performance because additional information is exploited from the observed features $\mathbf{x}$. This advantage does not come for free, however, since the calculation of the expectation operation for GMMs with full covariance matrices implies a higher computational complexity.

It can easily be seen that the GMM-based algorithm from Sec. 6.8 is a special case of Eqn. 6.83, if only a single state $N_{\mathcal{S}} = 1$ is employed in the HMM. In this case, the sum in Eqn. 6.83 degenerates to the conditional expectation $\tilde{\mathbf{y}}(m) = E\{\mathbf{y}(m)|\mathcal{S}_1(m), \mathbf{x}(m)\}$ because $P(\mathcal{S}_1(m)|\mathbf{X}(m)) = 1$. This conditional expectation is identical to the estimation rule (Eqn. 6.59) in Sec. 6.8.

There are also certain parallels of special cases of variant II of the MMSE rule to the linear mapping and piecewise-linear mapping methods from Sec. 6.7: if there is only a single state in the HMM and one Gaussian in the mixture model of the emission probability, that is, if $N_{\mathcal{S}} = 1$ and $L = 1$, the MMSE rule (Eqn. 6.83) also leads to estimation of $\tilde{\mathbf{y}}$ by linear transformation, though with consideration of the mean vectors of $\mathbf{x}$ and $\mathbf{y}$ (cf. Eqn. 6.59). With one Gaussian in the emission PDF models $\tilde{p}(\mathbf{x}, \mathbf{y}|\mathcal{S}_i)$ and more than one state in the HMM, that is, if $L = 1$ and $N_{\mathcal{S}} > 1$, the MMSE rule of Eqn. 6.83 resembles a piecewise-linear mapping approach with soft decision.

## 6.10 DISCUSSION

In the past, bandwidth extension algorithms for speech have reached a stable baseline quality: the artificial wideband output of a BWE system is in general preferred to narrowband telephone speech. Nevertheless, the quality of the enhanced speech is far from reaching the quality of the original wideband speech. It would be desirable to further improve the subjective speech quality of BWE systems.

With respect to the performance of wideband spectral envelope estimation, comparison of the theoretical performance bound (e.g. Fig. 6.19) with the best actually achieved estimation results (e.g. Fig. 6.20) yields a performance gap of about 3.2 and 2.3 dB for high-frequency (3.4–7 kHz) and low-frequency (50–300 Hz) BWE of telephone speech, respectively (Jax [128]). It is unclear, unfortunately, whether this gap can be closed by more sophisticated estimation schemes, because the theoretical bound, in general, is not tight. Some authors have come to the conclusion that improving the objective performance, as, for example, measured in terms of log spectral distortion, of (memoryless) BWE for speech may be very intricate (Epps [65], Nilsson *et al.* [186]).

To date, BWE for speech has mostly been developed for clean input speech. The vast majority of the published approaches do not consider any adverse conditions such as additive background noise or distortion of the narrowband input signal. To improve the acceptance in practice in the wide range of possible applications, the robustness of BWE for speech schemes has to be increased. Important issues in this respect are robustness against additive background noises, and against input signals that differ from the model assumptions, like music and so on. In such circumstances, at least the bandwidth extension system should switch to a secure fallback mode, for example, similar to one of the generic BWE algorithms as described in Chapters 3 and 5.