This chapter focuses on the cochleogram: a continuous function $r(s,t)$ of place $s$ and time $t$ reflecting the energy of basilar membrane segments. A generalization of the cochleogram, which includes periodicity as a third dimension will be addressed in the next chapter. Auditory modeling for speech signal processing focuses often on the nonlinear behavior of a filterbank description of the basilar membrane (Brown 1994, Patterson 1995, Tchorz 1999). As discussed in section 2.2, for the time being a linear system is preferred because it facilitates the separation of sound sources.

Section 3.1 addresses the implementation of the BM model. Section 3.2 studies spatial differentiation of the BM and the trade-off between frequency selectivity and group delay. The place-frequency relation of the BM and the response to sinusoidal stimuli is addressed in section 3.3. Section 3.4 addresses ways to estimate the contribution (in terms of energy) of individual harmonics of a known fundamental. The origin of stable ridges and the way ridges are estimated are discussed in section 3.5. The chapter is concluded with section 3.6 which addresses the cochleogram reconstruction technique that was used in section 2.12.
The Basilar Membrane Response

3.1 The Basilar Membrane Model

The basilar membrane (BM) model is based on the work of Duifhuis (1985), van Netten, Hoogstraaten and van Hengel (1997) who developed and validated a one-dimensional model of the human cochlea and implemented the model in FORTRAN. Although the model represents the middle ear as well as the cochlear partition (which includes the BM), it is referred to as the basilar membrane model. An overview of the basic structure of the model and its relation with the natural basilar membrane is given in figure 3.1. The model is developed for auditory research purposes and is optimized to be as physiologically plausible as possible with a one-dimensional basilar membrane model. The BM model was originally implemented as a nonlinear transmission line model. Special attention was paid to the numerical stability of the implementation. The original numerical model consists of 400 segments and functions at an internal sampling frequency of 400 kHz.

Figure 3.1. A schematic depiction of the basilar membrane (BM) and an electrical equivalent network of the BM model. The BM is modeled as a cascade of 400 second-order filters that model the cochlear fluid mass ($L_{\text{fluid}}$), the BM-segment mass ($L_{BM}$), the friction (damping) due to the basilar membrane movements ($R_{BM}$) and the stiffness of the BM ($C_{BM}$). The velocity of the natural basilar membrane is transduced to graded-potentials by approximately 3000 hair-cells and transmitted in the form of action-potentials to the brainstem by 30000 spiking neurons.
The main drawback of this model for speech recognition research is the high computational demand. In 1993, with a 33 MHz machine, the system was a factor 1000 slower than real-time. To improve processing speed above the rate provided by improving computer technology, a faster implementation of the algorithm was required. To benefit optimally from linear signal-processing theory, a linear version of the model is used. This choice is mathematically convenient and could be motivated by the argument given in section 2.2. It is an open question whether or not a biophysically more realistic nonlinear model might eventually be superior for ASR purposes.

An overlap-and-add filter bank implementation is chosen because it requires a minimal number of floating-point operations. The impulse response that defines the filter bank is based on the velocity of a linear 400 segment version of the model with an internal sampling frequency of 400 kHz. Its impulse response is down-sampled to 20 kHz to reduce the number of operations per second and to allow the system to work with standard speech databases at 20 kHz. To reduce computation time, 300 segments, corresponding to frequencies below 6100 Hz, are used and the impulse response is limited to 50 ms. For all experiments only each third segment is used so that only 100 segments have to be computed. Although this reduces the computational load by a factor of three, it reduces spatial resolution and sensitivity just as well.

The impulse responses of all 300 low-frequency segments\(^1\) are depicted in figure 3.2. As in figure 2.6 all negative values are set to zero. Figure 3.2 shows amplitude information, while the figures in chapter 2 all depict information in the quadratic “energy” domain. Accordingly, the dynamic range compression is chosen as \(x^{0.30}\) and not \(x^{0.15}\) as in section 2.3. This leads to dynamic range compression effects comparable to the nonlinear behavior of the intact cochlea. The dynamic range compression is only applied for visualization purposes and to compute the final output; prior to visual presentation all processing is linear.

The impulse response of figure 3.2 demonstrates the most important properties of the BM model. First of all, the figure is continuous in both time and place: it is possible to choose arbitrary paths through the impulse response plane of the BM without encountering discontinuities. The place-

---

\(^1\) Because it is often convenient to have a frequency-axis and a segment numbering that correspond closely, the segment numbering starts from the low-frequency side. This is contrary to the convention in auditory modeling.
The Basilar Membrane Response

continuity preserving signal processing

Frequency relation is visible as well: segments sensitive to the highest frequencies show many oscillations in a given interval, while the low-frequency segments show only a few. The impulse response profile shows why the excitation of the BM is often described as a traveling wave. Consider for example the prominent band that ends around $t=30$ ms at segment 1. This band can be interpreted as the top of a wave that starts at the high-frequency side and travels within 30 ms to the low-frequency side. This entails that the low-frequency side shows a delay compared to the high-frequency side.

A few examples of impulse responses from different segments are presented in figure 3.3. Each impulse response shows a delay, that decreases with increasing segment number. Each segment has a frequency to which it is most sensitive: its characteristic frequency $f_c$. This frequency corresponds to the most prominent frequency in the impulse response. The moment the envelope of the oscillation reaches a maximum can be interpreted as a delay associated

---

Figure 3.2. The positive values of the impulse response of the linearized basilar membrane model shows rapid oscillations and a narrow temporal envelope for segments sensitive to high frequencies. On the low-frequency side the oscillations are slow and the envelope is broad. The part of the impulse response that extents beyond 50 ms is discarded. Note that the strong band, constituting the second peak at each segment, can be described as the peak of a traveling wave. The very weak and broad vertical bands and the structure in the lower left corner originate in the original model from reflections from the apex.
The Basilar Membrane Model

Continuity Preserving Signal Processing

87

Notice that there is no true delay: the whole BM responds immediately due to the incompressibility of the cochlear fluid. A formalization of this “delay” will be studied in the next section.

The impulse response defines the filter function of the filter bank version of the basilar membrane model. The filtering is implemented as an overlap-and-add filter bank (Lynn 1998). Typically, the signal is filtered in blocks of \(2^{12} = 4096\) points. The impulse response is reduced from 1000 samples (= 50 ms) to 997 samples. This entails that 4096 - 997 + 1 = 3100 samples of the completely filtered BM-response become available after each block. The 3100 samples correspond to 155 ms (i.e., exactly 31 frames of 5 ms). For each segment the last 600 ms of the BM-response are stored in a buffer. The BM filtering is one of the two processing bottle-necks, so considerable effort was made to ensure that the Matlab implementation \(^2\) was as efficient as possible. The 100 segment BM is 10 times slower than real-time on a 400 MHz computer.

Figure 3.3. Examples of impulse responses of different segments (300-segment version). After a delay, that decreases with segment number, oscillations start. The average period of the oscillation corresponds closely to the characteristic frequency of the segment. The peak of the oscillation’s envelope correlates with the delay due to propagation effects. The first oscillations are the strongest. This corresponds with figure 3.2. The difference between the impulse response of segment 50 and 51 is very small at each point in time. This is a direct consequence of the physical coherence of the BM that cannot be guaranteed in generic filter banks.

with the segment. Notice that there is no true delay: the whole BM responds immediately due to the incompressibility of the cochlear fluid. A formalization of this “delay” will be studied in the next section.

The impulse response defines the filter function of the filter bank version of the basilar membrane model. The filtering is implemented as an overlap-and-add filter bank (Lynn 1998). Typically, the signal is filtered in blocks of \(2^{12} = 4096\) points. The impulse response is reduced from 1000 samples (= 50 ms) to 997 samples. This entails that 4096 - 997 + 1 = 3100 samples of the completely filtered BM-response become available after each block. The 3100 samples correspond to 155 ms (i.e., exactly 31 frames of 5 ms). For each segment the last 600 ms of the BM-response are stored in a buffer. The BM filtering is one of the two processing bottle-necks, so considerable effort was made to ensure that the Matlab implementation\(^2\) was as efficient as possible. The 100 segment BM is 10 times slower than real-time on a 400 MHz computer.
3.2 Sensitivity Versus Group Delay

Although implemented as a regular filter bank, the BM model behaves like a linear transmission line. Compared to more conventional cochlear filter banks, which are often based on gamma-tone filters (Brown 1994, Patterson 1995, Tchorz 1999), a transmission line shows a well-defined spatial continuity. And in contrast to frame-based methods, like FFT or LPC, it shows spatial as well as temporal continuity. To study the merits of spatial continuity the first and second order spatial derivatives of the output were computed with:

\[ x^0_s(t) = x_s(t) \]
\[ x^1_s(t) = x_s(t) - x_{s-1}(t) \]
\[ x^2_s(t) = x_{s-1}(t) - 2x_s(t) + x_{s+1}(t) \]

The superscripted index denotes the order of differentiation. These are used as input for the cochleogram estimation according to:

\[ r^i_s(t) = r^i_s(t - \Delta t)e^{-\frac{\delta t}{2}} + x^i_s(t)x^i_s(t) \]

Versions of the cochleogram of the word /NUL/, based on different orders of spatial differentiation, are depicted in the upper panels of figure 3.4. The differences are obvious: spatial differentiation highlights the differences between neighboring segments, which in turn leads to enhanced frequency resolution. The lower panel shows the cochleogram cross-section at \( t=175 \) ms. These figures show that the original basilar membrane model can hardly resolve individual harmonics: the first harmonic dominates the response of all segments. The first and especially the second order differentiated cochleogram represent individual harmonics much better because the segment’s response curves are sharper.

The increased discriminability comes at a prize: sharper response curves represent more specific frequency information, which in turn requires more temporal information. Consequently the response of the segments will be slower. The response time of a linear system can be formalized as group delay.

---

2. All presented work is implemented in Matlab. Matlab does not optimize speed, but allows rapid prototyping in combination with extensive visualization capabilities.
The group delay of a linear system can be defined as the center of gravity of the squared impulse response $h_s(t)$:

$$d_s = \frac{\int t |h_s(t)|^2 dt}{\int |h_s(t)|^2 dt}$$

Figure 3.5 presents the group delay of different segments for different orders of spatial differentiation. The lower $x$-axis shows the corresponding characteristic frequency of the segments, the upper $x$-axis shows a normalized position axis where the last segment corresponds to 100 (the number of segments that is used in most experiments).

The first 10% of the basilar membrane is sensitive to the narrow frequency range between 28 and 116 Hz and the corresponding group delay is long. The

Figure 3.4. Spatial differentiation sharpens the response of the basilar membrane segments. The cochleogram based on the 0th-order (original) signal is featureless. Differentiation highlights the differences between neighboring segments which leads to an enhancement of spectral detail. A related effect is that differentiation functions as first-order high-pass filter. This is noticeable in the lower panel as a relative enhancement of high-frequency segments that leads to a reduction of the average slope. The responses of the first and second order derivatives are scaled to match the height of the peak of the original response.
The Basilar Membrane Response

last 10% of the basilar membrane segments on the other hand represents a 2000 Hz broad frequency range, the corresponding group delay is short. Consequently spatial differentiation sharpens the response, but leads to an increased group delay. As a general rule, a segment’s group delay \( d \) is at least a few times the characteristic period \( T_c \) (the inverse of the characteristic frequency \( f_c \)) of the segment:

\[
d = cT_c = \frac{c}{f_c}
\]

(3.4)

The higher the multiplication factor \( c \), the narrower its response and the more discriminative the segment\(^3\). The group delay curves of figure 3.5 show some odd details in the first 20 segments. These are attributed to the limitation of the impulse response to 50 ms. The true impulse response of low-frequency segments extends well beyond 50 ms. This suboptimal choice is a compromise

\(^3\) The factor \( c \) is closely related to the quality factor \( Q \), defined in the frequency domain as \( f_c / \Delta f \), with \( \Delta f \) a measure of the width of the response curve.

Figure 3.5. Group delay as a function of characteristic frequency and segment number (upper axis). Group delay depends on the specificity of the cochlear filters. High-frequency filters are broadly tuned and show a short group delay. Spatial differentiation sharpens the cochlear filters and increases group delay. The shoulders and other effects on the low-frequency side of the BM stem from limiting the impulse response to 50 ms.
to optimize processing speed. Fortunately these low frequencies are not of central importance for speech and the cut-off at 50 ms is partially masked by the smoothing effects of leaky integration.

The most important effects of group delay occur when the pitch of a signal changes rapidly. The high-frequency segments follow the instantaneous pitch rapidly, while the low-frequency segments still reflect the instantaneous pitch of \( t \) minus 20 ms and more (see figure 3.5). This complicates virtually all aspects of further processing.

The place-frequency relation, as well as the trade-off between group delay and the sharpness of the response, the quality factor \( Q \), can be adjusted in the original BM model (Duifhuis 1985). This work is based on a version of the model which operates with a constant \( Q \) of 10. This gives the rather blunt response profile of figure 3.4 (blue line). For our analysis we use the second spatial derivative, the increased group delay has to be dealt with appropriately. This will be discussed in several contexts in the next chapters (in particular in the sections 4.6 and 5.1). The effect of spatial differentiation can be incorporated in the impulse response by computing the second spatial derivative of the original 300-segment impulse response. Selecting the impulse responses of each third segment leads again to a 100 segment model.

### 3.3 Place-Frequency Relation

Most figures of this work use a place-frequency relation that has been estimated experimentally. In the original BM model the resonance frequency of the uncoupled segments is chosen according to the Greenwood place-frequency relation (Greenwood 1990, 1991) were \( x \) is measured in mm from the apex:

\[
 f_{\text{res}}[\text{Hz}] = 190 \cdot 10^{0.6x[\text{mm}]} - 145 \text{ [Hz]} \quad (3.5)
\]

The segment index \( s \) can be related to \( x \) by correcting for the length of the BM (35 mm), the number or segments in the original model (400) and the fact that only one out of three segments are actually used:

\[
 s = \frac{400 \times [\text{mm}]}{3 \times 35[\text{mm}]} \quad (3.6)
\]
The interaction between segments, shifts the actual resonance frequency to a slightly lower value than the hard-coded $f_{\text{res}}$ of equation 3.5. Figure 3.6 shows the Greenwood place-frequency relation as the dashed line. The local resonance frequency defines the characteristic frequency $f_c$ and is depicted as the thin black line. It is always below the Greenwood place-frequency relation. Analogous to the characteristic frequency of a segment, every frequency has a characteristic segment. The actual place-frequency relation is based on measuring the response strength of each segment to a range of logarithmically spaced sinusoids of unit amplitude. The resulting matrix, of which the values are color-coded logarithmically, forms the background of figure 3.6. High-frequency segments respond stronger to low-frequency stimulation than vice-versa.
Since (quasi-)periodic signals consist of combinations of harmonics, it is useful to study the response of the BM to single frequency stimuli. The response of the BM to a certain fixed frequency is termed a sine response. Several examples are depicted in figure 3.7. Irrespective the driving frequency, all BM-responses have a similar asymmetrical form with a more prominent tail towards the high-frequency side than towards the low-frequency side. The figure depicts steady-state situations that are only reached after a sufficient number (e.g., ten) of oscillations and/or a few (e.g., five) times the integration time-constant $\tau$. Natural signals rarely show signal components that change slowly enough to fully justify this steady-state assumption. On the low-frequency side of the BM, the pitch as well as the amplitude are seldom constant enough during the 50 ms or more that are required to reach a steady-state. This results in broader responses than the ideal sine-response.

Figure 3.7. The BM-response to sine-wave stimuli of selected frequencies (in Hz). The upper axis provides the characteristic frequency (in Hz) of the segments of the lower axis. The red line shows the response of all segments of the BM to a unit-strength sinusoidal stimulus of 300 Hz. This frequency excites segment 25 most. This excitation drops much more steeply to the low-frequency side than towards the high-frequency side. Consequently, low-frequency stimuli mask high-frequency stimuli more than vice versa. The information in this figure corresponds to horizontal cross-sections of figure 3.6.

Since (quasi-)periodic signals consist of combinations of harmonics, it is useful to study the response of the BM to single frequency stimuli. The response of the BM to a certain fixed frequency is termed a sine response. Several examples are depicted in figure 3.7. Irrespective the driving frequency, all BM-responses have a similar asymmetrical form with a more prominent tail towards the high-frequency side than towards the low-frequency side. The figure depicts steady-state situations that are only reached after a sufficient number (e.g., ten) of oscillations and/or a few (e.g., five) times the integration time-constant $\tau$. Natural signals rarely show signal components that change slowly enough to fully justify this steady-state assumption. On the low-frequency side of the BM, the pitch as well as the amplitude are seldom constant enough during the 50 ms or more that are required to reach a steady-state. This results in broader responses than the ideal sine-response.
On the high-frequency side of the BM, steady-state is reached quicker, but random pitch fluctuations of natural signals broaden the responses as well. Nevertheless the sine-response is a very useful approximation that is used extensively in section 3.6.

**3.4 Estimating Individual Signal Components**

Since the BM model is linear, its response is a summation of the responses to the individual components of the driving sound events. In the case of a quasiperiodic sound event \( y(t) \) the input can be described as:

\[
y(t) = \sum_n a_n(t) h_n(t) \quad h_n(t) = \sin \left( \frac{2\pi n}{T(t)} t - \phi_n(t) \right)
\]

where \( a_n(t) \) denotes the amplitude of the harmonic contribution \( h_n(t) \). The harmonic is a function of the period contour \( T(t) \) and a phase function \( \phi_n(t) \).

The cochleogram of this signal is defined by equation 2.1. The square and a sufficiently long integration time-constant \( \tau \) ensure that the effect of the phase-term \( \phi_n(t) \) vanishes (except for some exceptional phenomena that are not considered here). In most cases \( a(t) \) changes slowly compared to the value of the time-constant \( \tau \). This means that \( a_n(t) \) can be treated (for short intervals) as a constant \(^6\) that scales the cochleogram contribution of \( h_n(t) \) with a factor \( <a_n^2(t)> \). The \( < > \) denotes temporal average as estimated by the leaky integration process. The cochleogram contribution of \( h_n(t) \) is denoted as \( R[h_n(t)] \). For arbitrarily developing harmonic contributions \( R[h_n(t)] \) is difficult to calculate exactly, but for slowly developing \( h_n(t) \), it can be approximated by the sine-responses as depicted in figure 3.6 and figure 3.7.

This means that the cochleogram \( R(t) \), resulting from a signal \( y(t) \) according to equation 3.7, can be approximated as:

\(^4\) This phase term might be estimated from the local basilar membrane movement.

\(^5\) The phase term can, in principle, counteract any effect of the period contour \( T(t) \) and might dominate the signal completely. These cases are not considered here.

\(^6\) This application of stationarity is justified since it applies to individual signal components.
Estimating Individual Signal Components

\[ R(t) = \sum_n a_n^2 R[h_n(t)] = \sum_n w_n(t) R_n(t) \]  

(3.8)

\( R_n(t) \) is the response of a unit-amplitude harmonic contribution \( h_n(t) \), approximated by a succession of the sine-responses of the characteristic segments corresponding to the temporal development of the local instantaneous frequency \( h_n(t) \). The weighting \( w_n(t) \) determines the scaling of this sine-response.

Any quasiperiodic signal can be approximated by superimposing correctly weighted response-curves for each harmonic contribution. The quality of the approximation is, as ever, a measure of the validity of the underlying assumptions. In general equation 3.8 can be applied safely on regions that are dominated by a single source: it can therefore be applied safely along ridges that result from periodic signal components (see the discussion about indications of periodicity at the end of section 4.3). The approximation of equation 3.8 will be used in section 3.6 (which addresses the reconstruction technique introduced in section 2.12) and is valid when the quasiperiodicity assumption is valid. Errors may occur during the onset when the BM is not yet quasi-periodic.

Normally, the weighting \( w_n(t) \) of the sine-responses is unknown and must be estimated from the signal. To estimate the contributions of the individual harmonics of the signal in figure 2.4, two different approaches are proposed. The first approach exploits the asymmetry in the sine-responses by neglecting masking towards the low-frequency side. In this case the signal in figure 2.4 is approximated by first weighting the sine-response corresponding to the frequency of the fundamental. This accounts for part of the excitation at the position of the second harmonic, the remainder is attributed to the second harmonic. At the position of the next harmonic, the contribution of all previous harmonics is subtracted and the remainder is attributed to the current harmonic. This process can continue until the frequency of the harmonics exceeds the characteristic frequency of the last segment, but in practice it is limited to BM regions where harmonics are resolved. This method works therefore particularly well for the first harmonics and will be used for cochleogram reconstruction in section 3.6.

7. When harmonics are not resolved it is often possible to skip harmonics so that each remaining harmonic approximates the best frequency of a single segment. The estimated contribution of the selected harmonics must then be shared with the harmonics that were skipped.
The second method is a numerical solution of the matrix equation: $Rw = E$. $E$ is the target cochleogram cross-section, $R$ the matrix of sine-responses associated with the frequencies of the individual harmonics and $w$ the vector with the desired weighting values. When applied to the signal in figure 2.4, the fundamental $f_0$ is $1/4.60=217$ Hz, as can be estimated from the TNC in figure 2.6. The associated harmonic frequencies are $nf_0$. The characteristic frequency of the last segment of the BM is 6100 Hz; the highest harmonic number that can be expressed is therefore $6100/217 = 28$. For each frequency a sine-response can be selected and added to the matrix $R$. Solving $w=R^{-1}E$ (in a least square sense) and setting negative values of $w_n$ to zero leads to the results in figure 3.8.

The upper panel depicts the target $E$ in blue, the lower panel presents the scaled contribution $w_n$ of each harmonic. The red curve gives the weighted sum of sine-responses. As can be seen the match is very good, which entails

![Figure 3.8](image-url)

Figure 3.8. Estimating and adding individual harmonic contributions. The blue cochleogram cross-section in the upper panel can be approximated by adding weighted harmonic contributions. The weighting of each harmonic $h_n$, depicted in the lower panel, can be computed by solving a matrix equation. The red line in the upper panel gives the weighted sum of 28 harmonics. The average mismatch for segment 15 and higher is less than 3%. The values of $h_n$ in the lower panel are scaled to provide a unit contribution of the strongest harmonic.
that the harmonic content of the first three formants were estimated correctly. The weights of the highest harmonics can only be estimated reliably around formant peaks. At other positions the sine-responses associated with the harmonics overlap almost completely and numerical errors will influence the results. Lower fundamental frequencies exacerbates this problem, but the use of a BM model with more segments alleviates the problem.

This is an efficient and rather elegant method for analyzing the harmonic content of a periodic signal when the fundamental frequency contour is known. The technique also works when the pitch of the signal changes within the validity range of the quasiperiodicity assumption (which seems to be the case for spontaneous speech). In this case, the approximation with the sine-responses is somewhat less valid and the effects of group delay effects have to be accounted for by choosing a set of frequencies that reflect the local instantaneous frequencies of the harmonics. Yet, this correction is straightforward if a correct pitch-contour is provided. Further work will extend this technique to deal with rapid onsets, its use on mixtures of periodic signals and its application to TAC-cochleograms.

3.5 Estimation of Ridges

As discussed in section 2.6, it is important to reduce the search space to allow computationally feasible implementations. The search space is reduced very efficiently by limiting TNC computation to ridges $s(t)$ in the cochleogram. The existence of more or less stable ridges is, again, a direct physical consequence of BM continuity. Two signal contributions can interact in different ways. An uninteresting case occurs when one signal contribution masks the other altogether. In this case, the weaker contribution cannot be estimated. More interesting cases arise when both exert a noticeable influence. An important situation occurs when both signal contributions have frequencies that correspond to a single segment or close neighbors. In this case, intervals with constructive and destructive interference alternate. This results in amplitude modulation, with a period corresponding to the inverse of the frequency difference between both signal components, and the formation of a ridge at the position corresponding to the weighted mean frequency of both components. The leaky integrated energy value associated with this ridge shows amplitude modulation. In noisy situations this may result in interrupted ridges. This type of interaction is important for harmonic
complexes at formant positions, which show amplitude modulations with the fundamental period of the signal. These modulations can be detected when the lowpass filtering has a sufficiently fine temporal resolution (which is generally not the case in the current implementation).

Another important interaction between signal components arises when the signal components correspond to segments that are further apart, so that both dominate their corresponding characteristic segment. Somewhere in between (due to the asymmetrical nature of masking usually close to the high-frequency segment) segments exist that feel a comparable influence from both components. These segments must follow two different frequencies without rupturing the BM. Consequently the local amplitude is reduced, especially when the signal components are anti-phasic. The corresponding local energy is small as well. *This leads inevitably to a situation with two energy peaks separated by a valley.* For signal contributions that persist for some time, the corresponding peaks string together to form temporal ridges. Ridge forming proves the existence of stable ridges corresponding to sufficiently separated continuously developing signal components. This is a transmission line property that cannot be guaranteed in general filter bank based BM models.

Ridge estimation is implemented straightforwardly. Candidate ridges are formed by stringing together peaks that differ less than 2 segments. In the odd case that ridges split or merge, the most energetic peak on the ambiguous side forms part of the continuation, the other peak forms the start of a new ridge or the end of an old ridge. When candidate ridges have a duration of at least 20 ms (4 frames) they are accepted as valid ridges. Ridge estimation can be improved by an approximation of the frequency and energy development of the ridges.

### 3.6 Cochleogram Reconstruction of TAC Selections

Section 2.12 provided an overview of the techniques that are used to transform the TAC-selections into a form that is suitable for HMM-based ASR purposes. This section studies the details of the reconstruction process that was used to remove the negative correlations in the TAC-selection. This reconstruction technique can also be applied to the more refined mask forming technique that will be developed in chapter 6.
Cochleogram reconstruction is a two-step process that is illustrated in figure 3.9. The first stage searches for evidence of resolved harmonics and uses this evidence to compute the harmonic contributions in the lower half of the reconstruction. The second stage adds information about the rest of the cochleogram using the mask and an approximation of actual activation patterns.

The first stage of the algorithm involves the estimation of coherent ridges in the (60) low-frequency segments. Ridges are formed, as in the previous section: with ridges longer than 15 ms accepted as candidates for harmonics. Since the fundamental period contour is known, it is possible to predict the segment numbers of the lowest harmonics. The ridges that are, on average, within 1 segment of the expected value of the first 5 harmonics are accepted as harmonics of the target signal. This criterion discards spurious ridges on the bases of a mismatch in temporal development. The number of harmonics that can be modeled this way depends on the spatial resolution of the basilar membrane. With a more sharply tuned BM model and a higher number of segments, a higher number of harmonics can be treated individually. In this case 5 harmonics were treated individually because the acceptance regions of the first 5 harmonics do not overlap in the present BM model.8 The algorithm is only weakly sensitive to the value of this parameter and works usually fine with up to 6 or 7 harmonics.

The upper left hand panel of figure 3.9 shows all candidate ridges. The energy development along these ridges is smoothed by replacing each value with a three point local average. The smoothed harmonic ridges are used to reconstruct an estimate of the original cochleogram by adding contributions of successive harmonics conform equation 3.8. This process is shown in the top panel of figure 3.10. The reconstruction starts with weighting the ideal sine-response (as shown in figure 3.7) of the fundamental. It is assumed that harmonics influence each other only upward in frequency (see section 3.4 for the justification of this assumption). At the position of the second harmonic, part of the energy can be attributed to the first harmonic, and the rest of the energy is used to weight the ideal sine-response of the second harmonic. In the top panel of figure 3.10, a large fraction of the energy of the position of the third harmonic must be attributed to the second harmonic, the fourth and fifth

---

8. This limits sufficiently resolved harmonics to the first 60 segments.
are relatively more important. The resulting partial reconstruction, based on 5 harmonics, is depicted in black.

The second stage of the algorithm is the reconstruction of the high-frequency range. Again the mask is used to pinpoint the regions that are most likely to represent information of the target. The selected values under the mask that exceed the partial reconstruction replace the values of the partial reconstruction. The result of this step is depicted in the lower left hand panel of figure 3.9. This stage leads to high-frequency contributions with unrealistically steep upward and downward slopes. The black peaks in the upper panel of figure 3.10 show this clearly. To make the reconstruction more realistic without adding extra information, the ridges of the mask can be augmented with flanks that represent masking effects consistent with a harmonic that excites the position of the peak associated with the flank. These can, again, be estimated from the sine-responses and added to the reconstruction. Finally the effect of leaky integration can be modeled as
exponential decay. The final reconstructions is shown in the lower right hand panel of figure 3.9 and drawn in black in the lower panel of figure 3.10.

Visual inspection shows that the reconstruction is often of high quality. The presented situation with a global SNR of 0 dB babble noise is close to the limits of the selection technique. In figure 3.10, the average instantaneous SNR per segment is -5 dB. Part of the signal, in particular the high-frequency range of figure 3.10, has a very unfavorable local signal-to-noise ratio. As can be seen in the lower panel of figure 3.10, a correct reconstruction is likely when the red target is close to the blue line that corresponds to the total energy. This corresponds to situations where the local SNR is favorable (SNR > 3 dB, red
stars) and the BM is dominated by the target. The difference between the red target energy and the black reconstruction shows that the energy of the fifth harmonic is underestimated, this is a result of a TAC estimation error due to the local optimization algorithm that will be discussed in section 4.6. The underestimation of the fifth harmonic was balanced by the selection of the two contributions at segments 81 and 88.

When the distance between the black and the blue line increases the probability increases that the reconstruction is incorrect. When the distance between the red and the blue lines corresponds to an SNR between 0 and 3 dB, the target can still dominate, but the influence of the noise is considerable. For negative SNR (no green stars), the BM segments are not dominated by the target and the reconstruction is likely to represent spurious contributions. Generally the quality of the reconstruction deteriorates gracefully. Selections will be reliable for local signal-to-noise ratio’s of 3 dB and better, and deteriorate rapidly with negative local signal-to-noise ratios. Below a local SNR of -3 dB, reconstruction leads to unreliable contributions. This provides a basis for the choice of a threshold of 0 dB of the background model in equation 6.2.

The link between the local SNR and the probability for a correct reconstruction (and consequently the assignment of information to the correct representation) relates the tuned autocorrelation to the experimental and theoretical work of Fletcher, French, Steinberg and Galt (French 1947, Allen 1993) that showed that the local SNR and not the spectrum determines the intelligibility of (nonsense) words.