A general sound recognition system (definition 1.4) must avoid the signal-in-noise-paradox. This chapter refers back to conclusion 1.12 that avoided the paradox by grouping and selecting evidence represented by coherent cochleogram areas into a representation on which auditory event (definition 1.11) formation can be based. The selected evidence can be used for sound resynthesis, cochleogram reconstruction, as input for automatic speech recognition systems, or as input for a more elaborate auditory scene analysis stage. As described in section 1.8, this work tries to identify signal components that are likely to have a favorable (local) SNR and to combine these into a single representation whenever they might stem from the same source.

Section 6.1 identifies the regions of the time-place plane of the cochleogram that are likely to have a high SNR; these areas are called auditory elements (conform Brown 1994). The section provides a selection method with a high preference for regions dominated by speech-like signals and a low preference for regions dominated by common noise-types. This addresses the first two steps of the measure of success in section 1.8. Section 6.2 quantifies the similarity between the selected information and the target signal by comparing the cochleogram of the target signal with a cochleogram based on resynthesized sound derived from information represented by the selected auditory elements. While section 6.2 addresses a measure of similarity based
on the shape of the cochleogram, the main task of this work is to select the most informative signal components of target signals in a wide range of signal-to-noise ratio's. This task is quantified in section 6.3 which addresses the third and last step of the measure of success in section 1.8.

6.1 Auditory Element Estimation and Selection

This section a number of simple physical thresholds select cochleogram regions. These place-time regions are registered as either accepted areas $A$ or as masks $M$. The accepted areas reflect the raw suprathreshold information. The application of a hard threshold may lead to adverse effects: when short intervals are discarded while their neighborhood is accepted, they are likely to be close to the threshold and consequently might be included to form coherent cochleogram regions. The masks provide such regions by combining segment contributions of a minimal duration $L$ in which holes of maximal duration $H$ are filled. Mask formation involves the application of a property of signals like speech: accepted signal contributions are limited to coherent contributions of a minimal length of 30 ms. This value corresponds to the duration of the shortest phonemes (e.g., /p/, /t/, /k/). This work uses a minimal duration $L = 30$ ms and a maximal hole size $H = 10$ ms.

It is assumed that a combination of constraints is more effective and more robust than the optimal use of a single constraint. Consequently the general approach of this section is to use a combination of low thresholds, rather than to use a single strict threshold (as applied in section 2.11 and section 4.3). Coherent area forming is based on:

- a model of the background energy
- the driving energy per segment
- a CPC-based measure of local dominance
- a TAC-based measure of compliance to a period contour
- combined criteria for periodic signal contributions
- combined criteria for aperiodic signal contributions
- combined criteria for speech-like signal contributions

The combined criteria are able to identify regions of the s-t-plane that are dominated by a single sound source. All thresholds are applied to a clean and to a noisy version (at 0 dB babble noise) of the standard target signal. As in
section 4.3 a click is inserted between the first and second word. Both cochleograms are depicted in figure 6.1. The remaining part of this section depicts raw selected areas $A$ as well as the corresponding mask $M$ for each threshold.

**Background model**

The first threshold is based on a model $C_B(s,t)$ of the background energy. Its main purpose is to discard cochleogram areas with a high probability to be dominated by the noise. A background model must be applied with utmost care, because it is crude and error-prone. Background models based on unchecked assumptions about the noise (like assuming a slow rate-of-change) are particularly dangerous. Ideally the background model must represent all signal contributions that do not vary synchronously with the speech (constant contributions and contributions that do not comply to the demands of the target class). Consequently, a perfect background model can only be estimated after the recognizable components of the signal have been classified (a direct consequence of the signal-in-noise-paradox).

The crude background model used here is based on the nonlinearly compressed energy (equation 2.3) because it represents perceptive effects better than the uncompressed energy (section 2.2). Typically, the background model is a function of time and place that is based on a moving average of the total energy with a large time-constant (e.g., $\tau_B>100$ ms). In this example, the onset of the target signal is too early to provide sufficient history. As an alternative, the background model is based on the average value $C_B^{\text{avg}}(s)$ and the associated standard deviation $C_B^{\text{std}}(s)$ of the nonlinearly compressed energy.
Auditory Element Estimation

\( r_s(t)^{0.15} \). It is assumed that the target dominates during period of voiced speech, consequently the average and the standard deviation are based on intervals where no pitch could be estimated:

\[
C_B^B(s, t) = \langle r_s(t)^{0.15} \rangle \\
C_B^{std}(s, t) = \langle \sqrt{\langle r_s(t)^{0.15} C_B^B(s, t) \rangle^2} \rangle
\]

The operator \( \langle \rangle \) denotes the average. The set \( P \) denotes the intervals in which a pitch could be estimated with the fundamental period estimation algorithm of section 5.1.

Another function of the background model is to discard quantization noise. The quantization noise level is determined by the dynamic range of the input signal, which in turn is determined by the type of signal, the dynamic range of the transmission channel, scaling of the data and/or the number of bits used per sample. In this example, it is set to \( c_{min}=1 \), which corresponds to 38 dB below the most energetic peak of the target cochleogram. (See figure 2.4 for an indication of this threshold in relation to the dynamic range of the speech signal.)

This leads to a background model \( C_B(s, t) \) with a lower limit of \( c_{min}=1 \) that is further defined as the average value of the nonlinearly scaled background \( C_B^B(s, t) \) plus a fraction \( c_{std} \) of the standard deviation \( C_B^{std}(s, t) \):

\[
C_B(s, t) = \max[C_B^B(s, t) + c_{std}(t) C_B^{std}(s, t), c_{min}]
\]

The criterion \( c_{std}(t) \) is set to \( c_{std} = 0 \) (\( v \) for voiced) for intervals with a TAC selection. This value is based on the observation at the end of section 3.6 that the TAC is likely to produce spurious contributions whenever the local SNR drops below 0 dB. Criterion \( c_{std}(t) \) is set to a less permissive \( c_{std} = 1 \) for intervals without TAC selections.

The application of the threshold \( C_B(s, t) \) to the cochleogram \( r_s(t) \), leads to an accepted area \( A_B \):

\[
A_B = \{ r_s(t)^{0.15} > C_B(s, t), \forall (s, t) \}
\]

and a corresponding mask \( M_B \):

\[
M_B = m_{L, H}(A_B) \quad L = 30 \text{ ms}, \quad H = 10 \text{ ms}
\]

1. The numerical noise level lies well below the quantization noise of the input signal.
The function $m_{L,H}$ is the mask forming function with parameters $L=30$ ms and $H=10$ ms. The effect of the background model is depicted in figure 6.2. The lefthand panels show the clean condition, the righthand panels show the noisy condition. The upper panels show the raw accepted regions, the lower panels show the masks that consist of segment contribution of at least 30 ms (with holes of 10 ms filled). Mask formation leads, in the noisy condition, to an area reduction of 9%. The use of a background model leads to an important reduction of the search space. The effect of different threshold settings for voiced and unvoiced regions is most noticeable in the low-frequency region.

Driving Energy and Decay
A second criterion is whether a cochleogram area is being excited or whether it reflects the decay of a response of a source that has effectively ended. Without driving energy, the cochleogram responses decrease with a rate determined by the lowpass filtering $L\{\}$, which, in this case, is implemented
Auditory Element Estimation

Figure 6.3. Cochleogram areas where decay dominates. The strong impulse at $t=500$ ms is the main cause of decay. Strong decay is not prominent in the noisy condition, but is of some significance in the clean condition.

as leaky integration with time-constant $\tau=10$ ms (equation 2.1). When a decay criterion\(^2\) $C_D(\tau)$ is based on $\tau_D=11$ ms, it can identify cochleogram areas $A_{\text{Decay}}$ with very little or no driving energy:

$$A_{\text{Decay}} = \left\{ \left( s, t \right) \middle| \frac{\partial}{\partial t} r_s(t) = \frac{\partial}{\partial t} \log r_s(t) < C_D(\tau) \right\}$$

(6.5)

Such areas are depicted in figure 6.3. In the noisy condition, decaying signal contributions are rapidly masked by noise, but in clean conditions areas dominated by decay may be of some importance and may help to identify pulses and offsets.

The acceptance criterion $\tau_D=11$ is set close to the limit of $\tau=10$ ms. This is a rather strict demand, but its complement, the areas $A_D$ with sufficient driving energy that we are interested in, is permissive. Much of the effect of the impulse, which was prominent in $A_{\text{Decay}}$, will be discarded in $A_D$.

\(^2\) In general, the decay criterion $C_D(s, t)$ may depend on segment dependent time-constants. The current implementation is based on a single global time-constant.
Then the next criterion is based on the characteristic period correlation and selects areas that are dominated by signal contributions that lead to the expression of their frequency at the corresponding characteristic segments (see section 3.3). This threshold has been introduced in section 4.3 as equation 4.10. In this section the local dominance criterion is renamed to $C_C(s)$ and relaxed with $C_C(s_{\text{high}})=c_{\text{high}}$ and $C_C(s_{\text{low}})=c_{\text{low}}$ and $c_{\text{high}}=0.02$ and $c_{\text{low}}=0.15$. The values for the intermediate segments are based on linear interpolation. The dominated area $A_C$ is defined by:

$$A_C = \left\{ (s, t) \left| \frac{r_f(t)}{r_i(t)} > 1 - C_C(s) \right. \right\}$$ (6.6)

This leads to figure 6.4. The figure uses a similar threshold for the high-frequency side as the lower panel of figure 4.7 (where $C_C(s_{\text{high}})=0.02$) and a threshold on the low-frequency side comparable to the middle panel of figure e...
Continuity Preserving Signal Processing

Figure 6.5. Accepted areas and associated masks for a TAC-selection based criterion $C_T(s) = c_T = 0.15$. Although the criterion is somewhat relaxed compared to the upper panel of figure 2.15 the application of a constraining period contour is still very powerful. It is the only signal property used for the grouping of signal components.

Compliance to a period contour

Compliance to a periodic contour is computed with the Tuned Autocorrelation. The upper panel of figure 2.15 showed selected values that represented 25% or more of the local signal energy. In general this criterion can be formulated as:

$$A_T = \left\{(s, t) \left| r_s \cdot T_0(t) \left( r_s(t) > C_P(s), \forall s, t \right) \right. \right\}$$ (6.7)
In this section the constraint is relaxed to $C_P(s)=c_P=0.15$. The resulting accepted areas and their associated masks are depicted in Figure 6.5. This TAC-based criterion is very powerful and it is the only (implemented) signal property that can combine different cochleogram regions in a single representation.

The rest of this section uses combinations of acceptance areas to form auditory elements.

**Periodic Signal Contributions**

Periodic signal contributions can be defined as signal contributions that dominate BM ranges by supplying sufficient driving energy and comply to a period contour. This leads to a combined criterion for periodic signal contributions and to a new accepted area $A_P$ defined as:

$$A_P = A_D \cap A_C \cap A_T \cap A_B$$

that reflects signal contributions defined by:

- $A_D$: sufficient driving energy
- $A_C$: sufficient local dominance
- $A_T$: compliance to a period contour
- $A_B$: exceeding the background model

To prevent adverse effects due to the multiple application of the gap-filling mask forming function $m_{L,H}$, the corresponding mask is defined as:

$$M_P = m_{L,H}(A_P)$$

and not as $M_P = m_{L,H}(M_D \cap M_C \cap M_T \cap M_B)$.

The resulting regions are depicted in Figure 6.6. The combination of constraints leads to a considerable reduction in the area of both $A_C$ and $A_T$ and a strong focus on the main signal contributions (i.e., strong individual harmonics and formants). This is not only the case for the clean condition, but also for the noisy condition. Inspection of the form and energy development of individual coherent areas in the noisy condition shows that they are separated into regions that tend to be dominated by either the background or the target signal. In the latter case the selected areas do often resemble the clean condition. This is an important step towards auditory event estimation.
Figure 6.6. Periodic signal components according to the combined criterion for active domination of the BM, compliance to a period contour and energy exceeding the background model. Note that individual coherent areas of the mask in the noisy condition are likely to belong to either the background, or to the target signal, in which case they often resemble the clean condition. This justifies the term auditory elements.

The acceptance criterion $C_B(s,t)$ for the background model $A_B$ is relaxed during intervals with a TAC-selection. The TAC-selection uses periodicity as a very specific signal property that correlates with the target signal, but (on average) does not correlate with all other signal contributions. The requirements of mask formation in combination with the relaxed background criterion and the thresholds for TAC-acceptance and dominance result in an effective reduction of spurious contributions.

**Aperiodic Signal Contributions**

The complementary information based on an estimation of aperiodic signal contributions is defined somewhat more complex:

$$A_A = A_D \cap A_C \cap A_B \cap A_{-p} \cap A_N$$

(6.10)

It reflects signal contributions defined by:

- $A_D$: sufficient driving energy
Continuity Preserving Signal Processing 165

Figure 6.7. The aperiodic signal contributions are defined as contributions that dominate regions of the BM actively during periods where no pitch could be estimated, and of which the CPC does not exceed the energy too much (here a fraction $c_N=0.03$). Note that the /T/ of the third word is selected in all conditions.

- $A_C$: sufficient local dominance
- $A_B$: exceeding the background model
- $A_P$: frames without an estimated period contour
- $A_N$: a CPC to energy ratio sufficiently close to (but above) unity:

$$A_N = \left\{ \frac{r^c(s,t)}{r^s(t)} < 1 + C_A(s) \right\}$$

(6.11)

The last threshold $C_A(s)$ is included to allow a distinction between periodic contributions, of which the CPC value can exceed unity, and dominating aperiodic contributions for which the CPC-value is defined to be close to unity (equation 4.9). It is set to $C_A(s)=c_A=0.03$ for all segments. This combination of thresholds limits the accepted CPC-energy ratio’s to values below 1.03 and above 0.98 down to 0.78 with decreasing segment number.

3. This choice is suboptimal since it assumes that the target signal is either voiced or unvoiced. Speech can be a combination of both.
Auditory Element Estimation

Figure 6.7 provides the raw areas and the associated masks. Only a few regions meet the demands. The impulse is visible in the raw areas of both conditions. The /T/ of /TWEE/ meets the demands as well.

Speech-like Signal Contributions

A combination of the accepted periodic and the aperiodic contributions can be used to select areas that might be target speech:

\[ A_S = A_P \cup A_A \quad \text{and} \quad M_S = m_{L,H}(A_S) \]  

(6.12)

These are depicted in figure 6.8. These masks form the final output of the area selection algorithms.

Although this mask will be referred to as the speech mask, it does not necessarily reflect speech. In fact very little knowledge of speech has been applied and consequently the bias towards the speech class is minimal. The main demands are pitch-contours with a duration of at least 50 ms that lie between 75 and 400 Hz (due to the limitations of the period contour algorithm of section 5.1) and segment contributions of a minimal duration of 30 ms (due to mask forming). These demands are implementation dependent and can be

---

166 Continuity Preserving Signal Processing
relaxed if necessary. Other types of signals will be selected as long as they comply to these general demands. Additional signal properties (e.g., pitch variability) may be included to reduce the range of accepted signal types. It is the responsibility of the pattern recognition stage to determine how the selected evidence ought to be combined and interpreted (see e.g., section 7.2).

There has not been any optimization of the thresholds. Each threshold is set to be permissive. The selection of speech like contributions depends on the combination of knowledge sources and not on the optimization of each threshold individually. It will be demonstrated in the next section that this approach is applicable to different types and levels of noise. Table 6.1 summarizes all threshold values. These values will be used in the next sections on a range of noise-types and signal-to-noise ratios.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Function</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L = 30$ ms</td>
<td>Mask: minimal duration of segment contribution</td>
<td>Related to minimal duration of speech-like signal contributions</td>
</tr>
<tr>
<td>$H = 10$ ms</td>
<td>Mask: maximal hole size</td>
<td>Reduces effect of hard thresholds</td>
</tr>
<tr>
<td>$c_D = \tau_D = 11$ ms</td>
<td>Decay: accepted time-constant for energy loss</td>
<td>Criterion somewhat larger than leaky-integration time-constant $\tau$.</td>
</tr>
<tr>
<td>$c_{\text{min}} = 1$</td>
<td>Energy: minimal value</td>
<td>Prevents contributions due to quantization noise</td>
</tr>
<tr>
<td>$c_u^u_{\text{std}} = 1$</td>
<td>Background: fraction of std. deviation above average</td>
<td>For intervals where no period contour could be estimated (unvoiced intervals)</td>
</tr>
<tr>
<td>$c_s^u_{\text{std}} = 0$</td>
<td>Background: fraction of std. deviation above average</td>
<td>Weaker criterion than $c_u^u_{\text{std}}$, used during intervals with TAC-selection</td>
</tr>
<tr>
<td>$c_{\text{high}} = 0.02$</td>
<td>Dominance: acceptance crit. for high-frequency segments</td>
<td>$1 - c_{\text{high}}$ must be close to but less than unity for high-frequency segments</td>
</tr>
<tr>
<td>$c_{\text{low}} = 0.15$</td>
<td>Dominance: acceptance crit. for low-frequency segments</td>
<td>Criterion for low freq. segm. to accommodate amplitude variation in target signal</td>
</tr>
<tr>
<td>$c_P = 0.15$</td>
<td>Periodicity: fraction of energy explained</td>
<td>Weaker criterion than in section 2.11</td>
</tr>
<tr>
<td>$c_A = 0.03$</td>
<td>Aperiodic: deviation from unity of CPC to energy ratio</td>
<td>Can make a distinction between dominating aperiodic noise and harmonic contribution</td>
</tr>
</tbody>
</table>

Table 6.1. Overview of thresholds for area selection and mask forming.
6.2 The Robustness of Mask Forming

The resynthesis process, as described in section 2.11, is developed as a means to form a visual representation of the selected information by converting the information represented by the speech mask via a conversion to sound (Slaney 1994) back to a cochleogram representation. The advantage of this procedure is that masking effects are determined by the model itself and not approximated imperfectly (as for example in section 3.6). This section provides a quantitative comparison between the acoustic information represented by the masks and the target sound.

Since the basilar membrane model is implemented as a finite impulse response (FIR) filter, it is possible to invert the filtering by reversing the impulse response in time and by compensating for segment dependent frequency effects caused by the double use of the basilar membrane filter. This compensation is based on the sensitivity of the BM as function of place. When the output of all filters are added, all BM information represented by the basilar membrane oscillations is converted back to a close approximation of the input sound.

The cochleogram of the resynthesized standard target signal (without the extra pulse) is depicted in the upper panel of figure 6.9. The resynthesis is based on an all-pass mask (i.e., a mask that spans the whole place-time plane). When the resynthesis is based on the clean signal in combination with the mask as estimated in 0 dB babble noise, a very similar cochleogram results (middle panel). The main differences are the low values during intervals that were not included in the mask. If the resynthesis is based on the noisy signal and the mask estimated from the noisy signal (lower panel), the cochleogram remains similar, but spurious contributions are introduced.

To measure the efficiency of the auditory element forming and selection process, cochleograms based on resynthesized signals were computed for a range of different levels and types of noise. The masks were estimated with the criteria of table 6.1 at SNR's that range from 20 down to -20 dB in steps of 5 dB. Four different noise types were added: babble noise, white noise, car factory noise and speech noise (all derived from the NOISEX-92 database, Varga 1992). The noise selections were randomly chosen from the noise files and scaled so that the root-mean-squares of the speech signal and the noise
had the desired SNR. Eight different noisy signals were created for each of the 36 noise conditions.

For most broadband noises it is unlikely that (even with an improved estimation technique) the required period contours can be estimated reliably below an SNR of approximately -6 dB. This is the result of the breaking-up of ridges due to masking by noise peaks and an associated reduction of the reliability of the local frequency estimates. The combined effect leads to a rapid increase in the search space which eventually prevents the estimation of the correct period contour. The inability to estimate period contours below approximately -6 dB SNR restricts the mask forming technique to SNR’s above this level.

For normalization purposes, it is necessary to compute masks of noise signals that contain no evidence for the correct period contour. These mask must be based on similar period contours as the other noise conditions to exclude effects of period contour estimation differences and/or errors. This led to the
Auditory Element Estimation

decision to base all masks on the period contours as estimated from the 0 dB SNR babble noise condition (see Figure 2.13).

The similarity $s_{te}$ between a target vector $\mathbf{r}_t$ (the clean signal) and an estimate of the target $\mathbf{r}_e$ from a noise condition $\mathbf{r}_n$ is based on a normalized dot-product:

$$
\cos(\alpha_{t,e}) = \frac{\mathbf{r}_t \cdot \mathbf{r}_e}{\|\mathbf{r}_t\| \|\mathbf{r}_e\|} \quad (6.13)
$$

This measure is normalized so that $s_{te}=1$ when $\mathbf{r}_t$ and $\mathbf{r}_e$ point in the same direction and $s_{te}=0$ when $\mathbf{r}_e$ equals the estimate $\mathbf{r}_n$ in noise without the target, i.e., when $\cos(\alpha_{t,n}) = \langle \cos(\alpha_{t,e}) \rangle$. The baseline $\langle \cos(\alpha_{t,e}) \rangle$ is computed as the average for each noise type ($N=72$). This leads to the definition of $s_{te}$ as:

$$
s_{te} = \frac{\cos(\alpha_{t,e}) - \langle \cos(\alpha_{t,n}) \rangle}{1 - \langle \cos(\alpha_{t,n}) \rangle} \quad (6.14)
$$

The similarity measure computed using the nonlinearly compressed cochleograms (interpreted as vectors) and based on the average similarity of the 8 samples per combination of noise type and SNR. The four noise types are normalized individually.

Five different similarities were computed.

1. **Target cochleogram vs. noisy cochleogram.** This comparison provides a baseline similarity, reflecting the similarity between the target and the unprocessed noisy cochleogram.
2. **Cochleogram of the noise vs. noisy cochleogram.** This comparison provides the similarity between the unprocessed signal and the cochleogram of the noise without the target signal. It provides complementary information to the similarity 1 and is mainly included as a check for consistency.
3. **Target cochleogram vs. resynthesized selection.** This condition provides the similarity between the target cochleogram and the cochleogram of the resynthesized auditory elements.
4. **Target cochleogram vs. resynthesized selection, voiced only.** The resynthesized cochleogram and the target cochleogram differ during ‘silences’: the target cochleogram has a minimum energy level due to (quantization) noise, while the resynthesized cochleogram shows an exponential decay during silences. The effects of this qualitative difference are reduced when the similarity is based only on the frames where pitch is defined.
5. **Resynthesized target vs. resynthesized selection.** This provides the similarity between the cochleogram of the resynthesized target (based on speech masks estimated from the clean condition) and the cochleogram of the
The Robustness of Mask Forming

Continuity Preserving Signal Processing

resynthesized selection. Because these two signals have been subject to the same processing steps it is the best indication of the robustness of the auditory element selection technique.

The results (as averages over the four noise conditions) are summarized in figure 6.10 and in table 6.2. Figure 6.10 depicts the similarity 1 as the solid black line. Its converse (2) is depicted as the dashed black line. Both lines intersect at approximately 10 dB, with a similarity $\alpha = 45^\circ$. This point denotes the SNR where the noisy cochleogram resembles the target as much as it resembles the noise.

The dashed red line shows the third similarity. The degradation between 20 dB and 5 dB is less than 2%, but in the clean condition the similarity is 0.895. When the comparison is exclusively based on the voiced intervals, the similarity increases to 0.931. The solid red line gives the average similarity for resynthesized selection. Because these two signals have been subject to the same processing steps it is the best indication of the robustness of the auditory element selection technique.
Auditory Element Estimation

The voiced intervals (4). This fourth similarity lies, for SNR better than 5 dB, 3.5% above the dashed red line. The solid red line degrades only 0.6% for the SNR range between 20 and 10 dB and 3% between 20 and 5 dB, while the reference (1) degrades 20% and 33%, respectively.

The comparison of the resynthesis based on the clean condition with the resynthesis of the noisy condition (similarity 5) is depicted as the solid blue line. It lies above the other measures for -6 dB and better. It degrades 0.7% between the clean condition and 20 dB (unprocessed reference 9%), 6% for the range between 20 and 5 dB (reference 33%) and 12% between 20 and 0 dB (reference 50%).

The horizontal distance between the unprocessed reference and similarity 5 may serve as an indication of the improvement of the SNR due to auditory element selection. For 5 dB SNR this improvement equals 18 dB (assuming that the noiseless condition equals 30 dB). Alternatively: the degradation of the processed data at 5 dB equals the degradation of the unprocessed data at 23 dB SNR.

The improvement of the SNR is more or less constant over a large range of SNR’s: between 5 and -10 dB the improvement lies between 15 and 18 dB. Above 5 dB the improvement is limited by the choice of an SNR of 30 dB associated with the clean condition.

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>1 target coch. cochleogram</th>
<th>2 noise coch. cochleogram</th>
<th>3 target coch. resynth</th>
<th>4 target coch. resynth voiced</th>
<th>5 resynth target resynth</th>
</tr>
</thead>
<tbody>
<tr>
<td>clean</td>
<td>1</td>
<td>0</td>
<td>0.895</td>
<td>0.931</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>0.909</td>
<td>0.410</td>
<td>0.897</td>
<td>0.938</td>
<td>0.995</td>
</tr>
<tr>
<td>15</td>
<td>0.834</td>
<td>0.567</td>
<td>0.899</td>
<td>0.939</td>
<td>0.982</td>
</tr>
<tr>
<td>10</td>
<td>0.714</td>
<td>0.710</td>
<td>0.892</td>
<td>0.930</td>
<td>0.966</td>
</tr>
<tr>
<td>5</td>
<td>0.560</td>
<td>0.832</td>
<td>0.874</td>
<td>0.906</td>
<td>0.934</td>
</tr>
<tr>
<td>0</td>
<td>0.398</td>
<td>0.917</td>
<td>0.843</td>
<td>0.858</td>
<td>0.876</td>
</tr>
<tr>
<td>-5</td>
<td>0.245</td>
<td>0.965</td>
<td>0.776</td>
<td>0.755</td>
<td>0.784</td>
</tr>
<tr>
<td>-10</td>
<td>0.129</td>
<td>0.987</td>
<td>0.664</td>
<td>0.557</td>
<td>0.640</td>
</tr>
<tr>
<td>-15</td>
<td>0.057</td>
<td>0.996</td>
<td>0.489</td>
<td>0.249</td>
<td>0.425</td>
</tr>
<tr>
<td>-20</td>
<td>0.027</td>
<td>0.999</td>
<td>0.392</td>
<td>0.086</td>
<td>0.326</td>
</tr>
<tr>
<td>noise</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 6.2. Different similarity measures for different SNR.
Figure 6.11. Development of similarity between masked target cochleogram and the noisy resynthesized versions for different types of noise. Above 10 dB there is little difference between the types of noise. The degrading effect of white noise is relatively weak because it is an inefficient masker of low-frequency information. Speech noise is an inefficient masker of the peaks in the spectrum. The minimal degradation difference in speech and white noise between -15 and -20 dB is attributable to the few remaining harmonics that are particularly resilient to speech noise.

Figure 6.11 shows that the degradation differences between the different noise types is small for SNR above 10 dB. Because of their nonstationary character babble and car factory noise are degrade the signal relatively efficiently. The degrading effect of white noise is relatively weak at high noise levels because it is an inefficient masker of the low-frequency information that dominates the general form of the cochleogram. Speech noise is also a relatively inefficient masker because it is defined as aperiodic noise with a spectral envelope equal to the average spectrum of speech. This entails that strong formants are still able to dominate locally while the rest of the harmonics are masked.

The measure of similarity between the clean and the noisy cochleogram, as depicted in Figure 6.10, correlates with the performance of standard HMM-based ASR systems trained on similar data. Figure 6.12 shows a prominent correlation between the similarity measures of this section with the results of the speech recognition experiment of section 2.13. This shows that an HMM-
Based system reacts similarly to an increase in the "distance" between a reference pattern and a noisy input as an inner product based similarity measure. This validates the choice of the normalized inner product as an indication of the quality of the auditory element estimation and selection technique of section 6.1.

A final note about the perceptive quality of the resynthesized sound. The perceptive quality depends in the first place on the amount of background noise that is mixed with the signal. When a signal is reconstructed with an all-pass mask (i.e., a mask that spans the whole place-time plane) only a direct comparison reveals a minimal perceptual difference due to the absence of frequencies above 6200 Hz. It is difficult to tell which of the two is the original. When the resynthesis is based on the clean signal in combination with the mask as estimated in 0 dB babble noise the resynthesis sounds muffled, but still quite natural. This demonstrates that the noisy selection, on which the mask is based, includes the features that are perceptively most relevant, while major distortions are avoided.

4. These signals are available at: http://www.bcn.rug.nl/andringa
If the resynthesis is based on the noisy signal and the mask estimated from the a 0dB SNR signal, it is intelligible, but the naturalness of the resynthesis is reduced due to distortions and spurious contributions. This reduction is attributed to the reduced presence of the background noise which forces the auditory system to include distortions of the target signal to the target stream. Some distortions that are not included in the target stream are perceived as so-called ‘musical notes’ (see section 2.11).

6.3 Robustness of Auditory Element Estimation

The previous section measured the similarity between the clean target cochleogram and a resynthesized cochleogram with the BM excitation represented by a mask. However, section 1.8 described a measure of success as:

1. identifying and describing, in terms of the temporal development of frequency and energy, the signal components of a clean target signal that are likely to have a high SNR,
2. selecting target signal components and discarding non-target components from a noisy version of the signal and
3. determining that the selected signal components represent the same temporal development of frequency and energy as the clean target.

The first and second tasks are performed by the auditory element estimation technique that was described in section 6.1. Each auditory element represents at least a single signal component with an associated energy and local frequency development (in the case of a quasi-periodic component a local instantaneous frequency development as well). Because section 6.1 combined auditory element estimation and selection, task 1 and 2 are combined in a single technique. What remains is the quantification of task 3.

An energy-domain development of a signal component comprises a description of the temporal development of frequency and energy (definition 1.20). The position and height of the associated ridge indicate its frequency and energy, respectively. Consequently, when a ridge has been preserved, a signal component has been preserved. The use of ridges prevents the dominance of a single strong harmonic and ensures a reasonable weighting of individual harmonics. The use of ridges entails that harmonics are counted
Auditory Element Estimation

Figure 6.13. Robust target mask and the intersection of the target mask with the set of ridges estimated from the clean signal.

Once. However, a focus on ridges may lead to an underestimation of the linguistic importance of harmonic complexes at formant positions that may give rise to only a single ridge.

It is possible to use the mask estimated from the clean condition (see the upper panel of figure 6.8) as the reference mask. But because of the absence of masking effects by other sources, the clean condition is rather special and not a reliable indicator of the robust signal components of a target signal. The clean mask it is therefore replaced by a mask based on cochleogram regions that were selected in 25% or more of arbitrarily chosen unmatched noise conditions above -10 dB SNR. This robust speech mask is depicted in the upper panel of figure 6.13. It is very similar to the mask derived from the clean condition, but approximately 10% of the low energy regions and approximately 15% of the low energy ridges are excluded. The energy development of the robust speech mask is based on the target cochleogram (clean condition). Resynthesized sounds based on the robust speech mask sounds a bit muffled, but perfectly intelligible and natural.

A measure of the overlap between a ‘lean’ ridge mask with a ‘fat’ (in terms of the broadness in the spatial direction of the coherent acceptance areas)
noisy speech mask, allows a robust comparison. Small estimation errors in either the ridge mask or the noisy speech mask are unlikely to lead to differences in the fraction of overlap. Because the ridge mask corresponds to a set of individual signal components (representing the most important linguistic information) it can be demonstrated that it is possible to select signal components that represent the same temporal development of frequency and energy as the clean target in a range of different types of background noises. This comparison will complete task 3 of this work.

The ridges in the robust speech mask are determined with the algorithm of section 2.6 and form a ridge template $M_R$ as depicted in the lower panel of figure 6.10. The ridges are only determined for periodic signal contributions, the aperiodic contribution (i.e., the signal component corresponding to the /T/) of the ridge mask is copied from the robust speech mask. The ridge template $M_R$ represents not only the most robust acoustic evidence of the target, it also represents pitch, formants, intonation and the main aperiodic contributions. Consequently the most relevant linguistic information is represented.\(^5\) When additional information about the width of formants (which can also be estimated from the noisy speech masks) is included, it is possible to reconstruct a synthetic cochleogram that resembles the original clean cochleogram with techniques as described in section 3.6. The robust and informative nature of the ridge mask will be used in a proposal for a robust recognition system in section 7.2.

Figure 6.14 shows an illustrative set of (very) noisy speech masks $M_S^n$ in red and the intersection with the ridge template, $M_R = M_S^n \cap M_R$, in black. A comparison between the lower panel of figure 6.13 and the black lines of the upper left hand panel shows that the most important ridges are selected in speech noise at an SNR of 0 dB (and above). This is true for the other noises, except for white noise that masks the high-frequency region efficiently. At -5 dB babble noise, most of the ridges are still conserved. At -10 dB white noise, the low-frequency ridges are still relatively unimpaired, but the high-frequency region is masked completely. The regions of the noisy speech masks of the other noise conditions reduce in area and begin to split up in smaller regions. Yet the ridge template still shows a considerable overlap. Below -10 dB the masks break down in a large number of small subregions.

\(^5\) This is an implicit validation of the use of conjecture 1.2.
Continuity Preserving Signal Processing

\[ \cos(\alpha) = \frac{M_R \cdot M'_R}{|M_R| |M'_R|} = \frac{M_R \cdot (M'_S \cap M_R)}{|M_R| |M'_S \cap M_R|} \quad (6.15) \]

These values are normalized according to equation 6.14. Table 6.3 summarizes the resulting similarities for the same set of noise conditions as were used in section 6.2. Figure 6.15 provides a graphical representation of the information in the table. The solid black line provides the unprocessed baseline similarity as in figure 6.10. The dashed black line denotes the average similarity over all noise conditions. The average degradation of the similarity between an SNR of 20 dB and 0 dB is 0.06 and can be attributed to the disappearance of the least energetic ridges. The most informative ridges, those that define the formant

Figure 6.14. Examples of area selection that illustrate the robustness of the match between the ridge template \(M_R\) and the area selection technique of section 6.1. The noisy mask \(M'_S\) is depicted in red, the ridges common to the noisy mask and the ridge mask are depicted in black. With decreasing SNR the masks reduce in area and the overlap with the ridge mask reduces as well. Yet even at -15 dB car factory noise, evidence of some important ridges remains in the noisy masks.
positions, remain almost unimpaired in this SNR range. Even at -10 dB, the average similarity is still 0.68. This is possible because the system has a priori knowledge of the period contour which allows the selection of a significant fraction of the strongest harmonics that dominate the similarity measure.

The improvement of the signal-to-noise ratio compared to the unprocessed condition (based on comparing cochleograms) as in section 6.2, is more than 20 dB (for 5 dB SNR or worse). There are some differences between the different noise types. Babble noise, car factory noise and speech noise behave similarly, white noise behaves somewhat aberrant because it has a flat spectrum while the other noise spectra behave more like $1/f$, which characterizes speech as well. This entails that white noise is an efficient masker of the high frequencies but an inefficient marker of low-frequency information.

The measure of distance in this section is related to a fundamentally different recognition approach than in section 6.2. That section was related to a traditional automatic speech recognition approach based on the comparison of an estimate of complete spectra (or spectral envelopes) with stored (stochastic) templates. Such an approach requires the transformation of a

---

Table 6.3. Similarities between the speech mask and the ridge mask for different noise types. The second column gives the unprocessed similarity degradation. The last column shows the average degradation of all noise conditions. The degradation at 10 dB SNR is still minimal. Babble and speech noise have almost identical degradation curves for low noise levels.
noisy input into a form with a reduced distance to the stored templates. The approach of this section is related to recognition systems in which word or syllable models use the estimated pitch to produce a detailed expectation (here represented by the ridge masks) that allows them to search (actively) for supporting and conflicting evidence. Such models are more versatile and more constraining than is usually implied by the term model in ASR research. In fact they behave much like the concept of a schema as used in the field of auditory scene analysis (Bregman 1990). Section 7.2 proposes a recognition system that is based on this type of actively searching models.

Figure 6.15. Similarity between the noisy mask and the target ridge mask. The black line denotes (as in figure 6.10) the unprocessed baseline similarity between the target cochleogram and the noisy cochleogram. The dashed black line gives the average similarity between $E_R^c = E_R^n \cap M_R$ and $E_R$. The colored lines provide the similarity for the different noise types. The difference between the unprocessed condition is approximately 25 dB (for 5 dB SNR or lower). The bar at -6 dB denotes the SNR below which it is unlikely (for most broadband noises) that a correct period contour can be estimated. The aberrant behavior of the white noise condition results because white noise is an efficient masker of high-frequency ridges, but an inefficient masker for low-frequency ridges. The curves of babble and speech noise overlap for positive SNR's.