Continuity preserving signal processing
Andringa, Tjeerd

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2002

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 05-02-2020
CHAPTER 2  

Introduction to CPSP

The previous chapter formulated the theoretical basis for the recognition of arbitrary signals on which the rest of this work is built. The purpose of this chapter is twofold. In the first place it provides a general overview of the basic **Continuity Preserving Signal Processing** (CPSP) techniques. Secondly, it describes a robust speech signal selection system and a first recognition experiment with a standard HMM-based speech recognition system.

This chapter focuses on the selection of quasiperiodic signals. In terms of time and energy, quasiperiodic signals form an important fraction of natural signals and are of central importance for speech recognition in noise. The last five numbered theorems of the previous sections form the basis on which the main line of thought is based. Operational and implementation details are presented in later chapters. The chapter starts with an overview of the most important techniques and representations that are addressed in this work.
2.1 Overview of Representations and Techniques

Figure 2.1 provides an overview of the techniques and representations discussed in chapters 2 to 6. Together, these techniques describe a new methodology to select coherent information about individual signal components from (possible) mixtures of sound sources (see sections 2.11 and 6.1). The selected information can either be used to resynthesize sound (see section 6.2) or it can be used to form a parametric description of the sound for coding and/or automatic recognition (see sections 2.12 and 3.6).

The input of the system is an arbitrary sound that is processed by a basilar membrane model that preserves continuity through time and place. The system identifies regions of the time-frequency plane (in this work actually a time-place...
plane) that have a very high probability of being dominated by a single source. A set of regions that is attributed to a target source is called a mask. The boxes in figure 2.1 near the arrow marked quasi-periodic signals are necessary to determine the fundamental period contours that are used to select periodic signals. The dashed boxes below the lower arrow marked all signals can be used to identify and characterize aperiodic signal contributions such as noise bursts and on- and offsets. This chapter introduces the techniques in the solid boxes; later chapters provide a more detailed treatment of all representations and techniques.

Section 2.2 addresses a continuity preserving split-up of the input sound into a large number of coupled frequency channels with a model of the basilar membrane (BM). A continuous representation of the temporal development of the energy as expressed by all basilar membrane segments is called a cochleogram. The cochleogram, defined in section 2.3, represents similar information as the short term Fourier energy spectrum, but the cochleogram is a representation of energy as function of time and place instead of time and frequency. Chapter 3 discusses the basilar membrane model and the cochleogram in depth.

To separate sound sources, knowledge about the desired target sources must be applied. Section 2.4 introduces a fundamental period contour (or pitch-contour) into the cochleogram definition to arrive at the Tuned Autocorrelation (TAC). A more detailed study of TAC estimation and the influences of interfering sounds is provided in section 4.6 and section 4.7 respectively. The tuned autocorrelation as well as the cochleogram form a subset of the Time Normalized Correlogram (TNC): a three dimensional quadratic domain representation of basilar membrane movements. Because the TNC conserves continuity though time, place (corresponding to frequency) and periodicity it forms the central representation of this thesis. The TNC is introduced in section 2.5. Chapter 4 is devoted entirely to the TNC and its properties.

For the tuned autocorrelation to be of practical use, fundamental period contours must be estimated from unknown input. This requires the identification of ridges in the cochleogram. Ridges correspond often to resolved harmonics and pinpoint regions with a high SNR and consequently indicate regions were reliable information can be derived from. Fortunately, these are also the regions where the quasi-stationarity assumption can be applied safely. Ridge estimation is treated in section 2.6, while section 3.5 explains how strong and resolved harmonics inevitably lead to ridges. When ridges represent resolved harmonics, it is possible to model the frequency
development of the signal contribution (harmonics) that produced the ridge. This leads to the estimation of Local Instantaneous Frequency (LIF) contours that are introduced in section 2.6 and described in more detail in section 4.5.

Local instantaneous frequency contours can be combined to arrive at the fundamental period contours required for tuned autocorrelation estimation. This is the subject of section 2.9 and chapter 5. The pitch estimation algorithms that are presented in this work are suboptimal in the sense that their basic assumptions are not suitable for a truly arbitrary input signal. The presented algorithms are therefore intended as a proof-of-concept and a basis for further development. The design decisions that lead to the presented system are outlined in section 2.8.

When suitable fundamental period contours have been estimated, they can be used to select quasi-periodic information with the tuned autocorrelation (section 2.10). The resulting TAC-selections can be processed further. As a first step it is useful to form a mask of the most reliable information. A simple example of a mask is introduced in section 2.11, while a more formal and complete treatment of mask forming is given in section 6.1, which addresses the formation of auditory elements, i.e., place-time regions with diverse properties that are likely to be dominated by a single source.

The information represented by the auditory elements can be used for the resynthesis of the target sounds (section 2.11 and section 6.2). Alternatively, it can be used to measure the individual signal contributions (e.g., section 3.4) and to reconstruct the clean (i.e., original) cochleogram (section 3.6). Or it can be used to parameterize the selected information (section 2.12) for automatic speech recognition experiments. Some preliminary ASR experiments are presented in section 2.13.

Chapter 2 does not address aperiodic signals. On- and offset effects, as well as the response of the TNC to an impulse and to white noise are treated in section 4.2. Since aperiodic signals represent a continuous distribution of frequencies (instead of the discrete set of periodic signals) one can actively search for continuous distributions of frequency components. This is discussed in section 4.3, which introduces the Characteristic Period Correlation (CPC) as a means to measure (local) dominance. The CPC and the cochleogram can be used as a basis for the identification and characterization of on- and offsets. On- and offset characterization is treated briefly in section 4.4.
Chapter 3 addresses the basilar membrane model but focuses on the main properties of the cochleogram as a continuous representation of energy as a function of basilar membrane position $s$ and time $t$. Chapter 4 addresses the Time Normalized Correlogram. The TNC $r(s,t,T)$ is a generalization of the cochleogram that includes periodicity $T$. Chapter 5 addresses the details of the fundamental period estimation techniques that were developed. Chapter 6 focuses on the identification and use of coherent signal contributions. These coherent signal contributions resemble the concept of auditory events that was introduced in chapter 1. Finally, chapter 7 summarizes the main findings and reflects on the whole work.

2.2 The Basilar Membrane Model

Figure 2.2 shows a very schematic representation of the essential features of the basilar membrane. The basilar membrane is a coiled-up structure with a length of 3.5 cm that is situated in the cochlea, a snail-house-like structure of about 1 cm$^3$. The side of the basilar membrane near the opening of the snail-

---

1. Note that energy is not referred to by the symbol $E$ as is customary. The representation where the energy measure is derived from is a correlation function for which the symbol $r$ is typically used.
house is most sensitive to frequencies of about 20 kHz. Further inside the cochlea the frequency to which each position is most sensitive decreases down to 20 Hz according to an (approximately) logarithmic place-frequency relation. The frequency-range of the basilar membrane is therefore 3 orders of magnitude or about 10 octaves. Approximately 3000 hair cells, evenly spread along the basilar membrane, transduct the local vibrations to graded-potentials, which in turn are coded as action-potentials and transmitted by 30000 neurons to the brainstem. The axons of these neurons form the auditory nerve. Figure 3.1 on page 84 shows some of these details.

This thesis uses a (simplified) linear, one-dimensional transmission line model of the basilar membrane (Duifhuis 1985, van Hengel 1996). The most relevant properties of the model are continuity in both time and place and a one-to-one place frequency-relation. This entails that the basilar membrane model can be interpreted as a filter bank with physically coupled filters: neighboring filters show similar displacements at all points in time.

Although the original basilar membrane model can be nonlinear like the actual basilar membrane, a linear version of the model is used. This allows an efficient implementation as an overlap-and-add filter bank and it helps to solve the central problem: how to separate a mixture of signals. After all, linearity entails additivity, which can be interpreted such that a mixture of signals $a$ and $b$ can be split without introducing cross-terms that depend on both $a$ and $b$. The absence of cross-terms, which cannot be guaranteed in most nonlinear system, simplifies the design and implementation of a sound separation system.

The original basilar membrane model requires an internal update frequency of 400 KHz and has 400 segments that span the full human frequency range. To reduce processing time the model was reimplemented as a filter bank with 100 channels spanning a frequency range between 30 and 6100 Hz. The filter bank implementation requires an in- and output sample-frequency of 20 kHz. This reduction of the number of channels trades spatial and temporal resolution, and indirectly noise robustness, for computational efficiency.

### 2.3 The Cochleogram

A time-frequency representation, like an FFT-based energy spectrogram is thought to represent the relevant information for the perception of speech and
can be interpreted straightforwardly. Unfortunately (given the discussion in section 1.7) it is discontinuous in both time and frequency. A spectrogram-like time-frequency representation, continuous in place (and indirectly frequency) can be computed by averaging the energy of (overlapping) frames of each basilar membrane segment. However this procedure implies quasi-stationarity, which (conform theorem 1.21) ought to be avoided since the input is not yet identified as a signal for which quasi-stationarity holds.

A continuous alternative in both time and place for the FFT spectrogram is the leaky integrated square of the displacement or velocity of the basilar membrane segments.\(^2\) The use of velocity is preferred over the use of displacement because the use of velocity enhances high-frequency components, which reduces the masking effects of high-frequency components by lower frequency components (see section 3.3). Leaky integration describes a process were the system, at each point in time, loses information about its previous state, but learns about the present. In this case:

\[
    r_s(t) = r_s(t - \Delta t)e^{\frac{-\Delta t}{\tau}} + x_s(t)x_s(t) \quad s = 1, ..., s_{\text{max}}
\]  

\(r_s(t)\) denotes the value of the leaky integrated energy of segment \(s\) at time \(t\), \(\Delta t\) is the sample period, \(t-\Delta t\) denotes the time of the previous sample, \(x_s(t)\) is the current output value of the channel. The time constant \(\tau\) of this first order system determines a scope of memory. For large values of \(\tau\) the value of exponential function is very close to unity, for small values the influence of the exponential function becomes more prominent since it reduces the contribution of the previous value of \(r_s(t)\). The square term \(x_s(t)x_s(t)\) is non-negative, hence \(r_s(t)\) is non-negative as well.

Equation 2.1 can be generalized to:

\[
    r_s(t) = L\{x_s(t)x_s(t)\}
\]  

where the operator \(L\{\}\) denotes any form of lowpass filtering.

The value of \(\tau\) is kept at 10 ms throughout this thesis. Real neurons perform a leaky integration process as well and 10 ms is a normal value for neurons (Weiss 1996, Segev 1995). While the input of equation 2.1 is the squared basilar

---

\(^2\) Note that the use of the basilar membrane displacement leads to a true energy representation. The use of the BM velocity leads to a representation that is related to energy via a frequency dependent correction factor. Although not completely justified, the term energy will be used throughout the work.
membrane velocity (Lim 1985), the neurophysiological equivalent is the all-positive amplitude compressed half-wave rectified basilar membrane velocity. The half-wave rectification is performed by the hair cells in the organ of Corti. The natural system shows a dynamic range compression of the BM movements $x$ that is often approximated as a cubed-root: $x^{1/3}$ (Stevens 1957, Hermanski 1990).

Dynamic range compression is necessary to bring all relevant features within the same range. This is important because $r_s(t)$, computed according to equation 2.1, has a dynamic range that, due to the nature of speech, can be 50 dB or more (Furui 1989). To compensate for the square in equation 2.1 the effect of the cubed-root is doubled and approximated by $x^{0.15}$:

$$R_s(t) = [r_s(t)]^{0.15} \quad (2.3)$$

All signal processing in this work is performed in the linear domain. Equation 2.3 is applied either as a last step of processing or prior to visual presentation (as is the case in most figures).

Since leaky integration is a low-pass filtering process, the output $r_s(t)$ can be downsamped to sampling-rates in the order of the integration time-constant. To accommodate sharp onsets a sampling-rate of 200 Hz, corresponding to 1 sample per 5 ms, is chosen. This leads to the cochleogram as the desired doubly continuous time-frequency representation. Figure 2.3. shows the cochleogram of the Dutch word /NUL/ (English: ZERO), spoken by a female speaker. This word is part of the target sentence /NUL EEN TWEE DRIE/ (English translation: /ZERO ONE TWO THREE/) that will be used throughout this thesis. Unlike the FFT-spectrum of figure 1.2, the cochleogram conserves continuity in time and place (or frequency).

It is instructive to consider the main structures of this figure. The broad red band, starting at approximately $t=50$ ms and $f=220$ Hz, is the first harmonic $h_1$ corresponding to the fundamental frequency $f_0$. The fundamental frequency rises during the utterance to a value above 350 Hz. The band above and parallel to the first harmonic is the second harmonic $h_2$. The lowest few harmonics form the first formant $F_1$. A second formant $F_2$ becomes visible after the transition from the /N/ to the /U/ at $t=120$ ms and drops during the

---

3. This is an application of quasi-stationarity. It is assumed that the local change in energy and frequency of a noisy signal can be modeled sufficiently well with a sampling-rate of 200 Hz. Speech-scientists often use 100 Hz as sampling-rate. The optimal sampling-rate may be a function of the basilar membrane position.
The Cochleogram

Figure 2.3. The cochleogram of the word /NUL/. Since it is produced by a single vocalization, it shows a single coherent development. The cochleogram is sampled each 5 ms. Notice the discontinuity at higher frequencies marking the end of /N/ and the onset of the /U/. The /L/ appears as a glide of the second formant between t=200 and t= 350 ms, that sequentially highlights the best matching underlying harmonic. The vertical line indicates the position where the information of figure 2.4 is derived from.

/L/ from 2000 Hz to about 900 Hertz. Notice that this change of formant position entails that different harmonics succeed each other as the most prominent local frequency contribution. A third formant $F_3$, is marginally visible during the /N/ but becomes prominent during the rest of the utterance. In the higher frequency regions a fourth and vaguely even a fifth formant are visible.

The transition from the /U/ to the /L/ is smooth, the transition from /N/ to the /U/ is partially discontinuous due to the transition from the nasal /N/ to the vowel /U/; at the end of the /N/ the tip of the tongue leaves the hard palate, allowing the oral cavity to resonate in addition to the nasal cavity. Notice that the onset discontinuity of the word is sharp and the offset is smooth. This is due to the exponential decay of the leaky integration process and the ringing-out effect of the basilar membrane in combination with the nonlinearity of equation 2.3.
A vertical cross-section of the cochleogram at $t=175$ ms is depicted in Figure 2.4. This figure shows a representation of the energy distribution as function of segment number (the lower horizontal axis) or the corresponding frequency (upper axis) corresponding to the information under the vertical line in Figure 2.3. Note the peaked structure. For low segment numbers the peaks correspond to resolved harmonics. For higher segment numbers the individual harmonics become less well resolved and merge eventually into formants. This behavior is a direct consequence of the nonlinear place-frequency relation. Several harmonics are depicted in the figure. The first three, the 9th, the 13th, the 18th and the 25th harmonics dominate the response. The 4th to 8th harmonics are just resolved, for the 10th to 12th harmonics only minimal visible evidence exists. These harmonics are (partially) masked by the other components. Although the higher harmonics are not resolved, they do contribute to the shape of the formants and contribute to the timbre of the vowel /u/.
Entrainment of segments by a dominant signal component is a very important property of a transmission line model and is due to the fact that the basilar membrane forms a single continuous structure. When a prominent signal component drives a certain segment, the segment will drag its neighbors along and they drag their neighbors along, etc. This effect attenuates rapidly as a function of place. Only the signal components that can overcome the recruitment-effect of other signal components will dominate locally and produce peaks. Entrainment is, like masking, more prominent on the high-frequency side, than on the low-frequency side. Entrainment and dominance is studied in more detail in section 4.3.

The first 12 segments show a low and irregular response with a value of around 0.7. This response is due to quantization noise. The highest value, just 7 segments away (in the original model 21 segments), is 3.25, a factor 4.7 higher. Considering the nonlinear compression in equation 2.3, this corresponds to a difference of approximately 90 dB. This is consistent with the dynamic range of 90 dB that is associated with the 16 bit input.

2.4 Tuned Autocorrelation

Splitting a mixture of signals without certainty about the signals origin requires the use of the weakest possible basic assumptions, i.e., the use of universally valid signal properties. An important general property is periodicity. In both speech and music, periodic signals represent the largest fraction of time and energy. Perfectly periodic signals do not occur often; most natural signals show amplitude and/or frequency modulations due to source properties.

A sound event $y(t)$ is quasiperiodic$^4$ with fundamental period contour $T(t) = 1/f_0(t)$, if for each harmonic $y_h(t)$:

$$y_i(t) = y_i(t + T(t))$$ (2.4)

If the harmonic $y_h(t)$ of the sound event entrains segment $s$ of the basilar membrane, the response $x_s(t)$ of the segment will show quasiperiodicity as well. Consequently:

$^4$ Although the amplitude of a quasiperiodic signal $y(t)$ can be a function of time as well, the effects of amplitude modulation are usually small for consecutive cycles and will be ignored in this section.
If $T(t)$ is known, equation 2.5 can be combined with equation 2.2 to yield:

\[
 r_{s,0}(t) = L\{ x_s(t)x_s(t) \} \\
 = L\{ x_s(t)x_s(t + T(t)) \} \\
 = r_{s,T(t)}(t) 
\]  

This entails that, under the condition that $T(t)$ is the correct fundamental period contour, $r_{s,T(t)}(t)$ closely approximates the cochleogram contributions $r_{s,0}(t)$ for all segments that are entrained by the sound event $y(t)$. This is important because $T(t)$ is a signal property with a very high probability of being unique for sound event $y(t)$. The set of values $r_{s,T(t)}(t)$ is defined as the Tuned Autocorrelation\(^5\) (TAC), because it is based on autocorrelation values $x_s(t)x_s(t + T(t))$ and tuned to a fundamental period contour $T(t)$ (and hence also to a fundamental frequency contour $f_0(t) = 1/T(t)$).

Equation 2.6 holds only for a correct fundamental period contour. For fundamental period contours that are not correlated with the contour of the target source, the values of $x_s(t)$ and $x_s(t + T(t))$ will not correlate and their average will be close to zero. This means that the TAC has values similar to the energy measure of the cochleogram for a correctly estimated period contour and values close to 0 for randomly chosen or uncorrelated period contours:

\[
 r_{s,T(t)}(t) = \begin{cases} 
 r_{s,0}(t) & \text{for correct } T(t) \\
 0 & \text{for uncorrelated } T(t) 
\end{cases} 
\]  

This property forms the basis for the assignment of information of quasiperiodic sound events into auditory events.

When it is not known which segments are entrained by the quasiperiodic source, the TAC of all segments is computed with:\(^6\)

\[
 r_{s,T(t)}(t) = r_{s,T(t)}(t - \Delta t) e^{-\frac{\Delta t}{\tau}} + x_s(t)x_s(t + T(t)) \\
 s = 1 \ldots s_{\text{max}} 
\]  

---

\(^5\) Papoulis 1984 defines the autocorrelation $R_{xx}$ of a real-valued random signal as $R_{xx}(t_1,t_2) = E[x(t_1)x(t_2)]$, where $E$ denotes an expectation operator. The TAC $r_{T(t)}(t)$ is a subset of the generic autocorrelation: $r_{T(t)}(t) = R_{xx}(t_1,t_2) = E[x(t_1)x(t_2)]$.

\(^6\) The actual formula (see equation 4.16) is slightly more complicated due to the effects of group delay. This phenomenon will be ignored in this overview, but it will play an important role in subsequent chapters.
Two examples are shown in Figure 2.5. The upper panels show the cochleogram of the word /NUL/ (conform Figure 2.3) and the positive values of the associated TAC. The lower panels show the cochleogram of this signal when cocktail party noise is added resulting in a signal-to-noise ratio of 0 dB (equality of signal and noise energy). The lower right hand panel shows the associated TAC. Compared with the panel above, most of the prominent structures are conserved, but some of the background has been “selected” as well. This cannot be avoided (see section 4.7). The TAC is not defined over the complete 500 ms, since the period contour of the sound event is only defined when the sound event is present. Note that negative values of the TAC representation are set to zero in visible representations only. This will be done throughout this thesis.
A tuned autocorrelation that results from a properly estimated period contour represents quasiperiodic information consistent with this contour. There is no guarantee that all information belongs to the same source, it is however guaranteed that all periodic contributions of the target source that entrain BM-regions will be represented.

The tuned autocorrelation is very robust. This has several reasons. A first reason is that the tuned autocorrelation selects all segment ranges dominated by target harmonics. In the case of broadband signals, like speech, in which a few harmonics or formants dominate, a peaked cochleogram results. The probability that formants (similar structures) of other sounds produce even stronger peaks that dominate the same regions even more prominently is usually small (but not zero). This probability is of course strongly dependent on the signal-to-noise ratio (SNR) and the distribution of energy over the frequency range. With common broadband signals that mask the target speech at a signal-to-noise ratio of -6 dB (ratio=1:4), the number of unmasked peaks of the target speech is reduced to a level where it becomes difficult to find a set of reliable starting points for the search of auditory events. Human speech perception deteriorates rapidly in these conditions (Plomp 1979, Alefs 1999).

A second reason for the robustness of the TAC is that a source does not need to dominate to provide a consistent local contribution. As long as the average contributions $x(t)x(t+T)$ of a less dominant source is larger than the average of $x'(t)x'(t+T)$ of a source that is dominating locally, the less dominant source will provide a positive contribution, even if it is masked optically. Since there are no coherent peaks, this situation does not provide reliable starting points for auditory event estimation. Yet this might explain why some noisy sentences cannot be perceived on first presentation when the listener does not know what to expect, whereas the same sentence is recognizable when the listener could form a correct expectation. For example, a naive listener might have difficulties with a target sentence at an SNR of -6 dB, while a experienced listener can perceive each word of the target sentence at -10 dB or less.

The most important problem with the application of the TAC is the necessity of a correct estimate of the fundamental period contour $T(t)$. Since it is not directly available, it has to be estimated from the signal. There exists an abundance of pitch estimation techniques, but none of these performs adequately on arbitrary (noisy) signals. The tuned autocorrelation is only a useful technique if a robust pitch estimation technique is available. Such a technique will be proposed in section 2.9. The next section addresses a
generalization of the TAC that is used as the theoretical basis for most of the techniques described in this work.

2.5 Time Normalized Correlogram

Equation 2.8 presents a subset of a more general continuous autocorrelation function:

\[ r_{x,T}(t) = L\{x_s(t)x_s(t+T)\} \quad s = 1, \ldots, s_{max} \quad T = [0, T_{max}] \]  

\( r_{x,T}(t) \) is typically implemented as a time-evolving matrix of dimensions (\# segments) x (\# periods) that is called the Time Normalized Correlogram (TNC). Of central importance is that the TNC is continuous in time, periodicity and place (with place related to frequency). The positive values of the TNC can be depicted in a similar way as the TAC-spectrograms. This is shown in figure 2.6. This figure shows the TNC for \( t = 175 \) ms in the middle of the /U/ of NUL. The vertical line at \( T = 0 \) corresponds to the energy spectrum that was depicted in figure 2.4. The vertical band at \( T = 4.6 \) ms represents the TAC for the fundamental period \( T_0 \). This band is repeated around 9.2 ms for \( 2T_0 \). These bands form the peaks of a large vertical structure that narrows as the frequencies of the individual harmonics increase. Each broadband quasiperiodic source has a similar structure; which is mainly determined by the instantaneous fundamental period.

The name TNC is derived from the fact that its definition in equation 2.9 ensures that whenever the first full period of a quasiperiodic signal ends at time \( t_0 \), its TNC starts to build-up at \( t_0 \) irrespective of the period \( T \) of the signal: for \( t < t_0 \) the temporal average of \( x(t)x(t+T(t)) \) is close to zero, while for \( t \geq t_0 \) the average is large and positive and independent of the value of \( T(t) \). This form of time-of-onset normalization (that does not hold for \( L\{x(t)x(t-T(t))\} \) ) helps to study the temporal development of all types of sources. A full discussion, where different definitions of correlograms are compared, is given in section 4.1.

Since it is unlikely that uncorrelated sources show a similar development of the instantaneous fundamental frequency, the probability is low that the vertical structures of different sources overlap. This is not the case for the energy term at \( T = 0 \) where the energy of all sound events get expressed on top
Continuity Preserving Signal Processing

Figure 2.6. The Time Normalized Correlogram derived from the /U/ of /NUL/ at \( t=175 \) ms. The TNC allows to follow arbitrary continuous paths through time (\( t \)), place (\( s \)) and periodicity (\( T \)). The vertical line at \( T=0 \) corresponds to the energy spectrum. The structure at \( T=4.6 \) ms is the TAC for the fundamental period \( T_0 \). This structure is mirrored around 9.2 ms for \( 2T_0 \). Note the way harmonics and formants are expressed.

of each other. The introduction of periodicity as an extra signal dimension allows not only a mixture of a periodic and an aperiodic signal to be split, but also mixtures of quasiperiodic signals! Note that this is partly an idealization: the combination of two or more quasiperiodic signals leads to a superposition of the individual TNC’s that is more difficult to interpret than a single one.\(^7\)

The vertical cross-section of the TNC corresponds to an autocorrelation lag \( T \) for all segments \( s \). The horizontal cross-section corresponds to the full running autocorrelation of a single segment. For aperiodic signals the correlation decreases as a function of \( T \), but since this source is periodic, the autocorrelation has the appearance of a cosine. Notice that most segments are dominated by a single harmonic. This is most prominent for segments that correspond to the lower harmonics. The periodicity of the local running

\(^7\) Interference effects between the different signals complicate the interpretation further.
autocorrelation reflects the frequency of the segment’s main driving force as a function of time. The first period that occurs in all segments is 4.60 ms which corresponds to 217 Hz. For the second harmonic the second period peaks at 4.60 ms. This corresponds to an instantaneous frequency of $1/(4.6/2)=434$ Hz, as to be expected. Just above 2000 Hz a region of the BM is dominated by the ninth harmonic. This region corresponds to the second formant. Note that the position of the tenth harmonic cannot be estimated as it is masked by the ninth. The third formant gets expressed just below 3000 Hz and is dominated by the 13th harmonic at: $(4.6 \text{ ms}/13)^{-1}=2.9$ kHz. Note that the TNC allows the estimation of instantaneous local frequencies with high accuracy. This is a direct consequence of the avoidance of a frame-based approach and the conservation of continuity. Section 4.5 provides the details of the local frequency estimation algorithm and discusses its accuracy.

The TNC is an extremely rich representation, but one of its most important features is that:

The TNC can represent arbitrary continuous paths through time $(t)$, place $(s)$ and periodicity $(T)$.

This means that if we know or hypothesize a period contour $T(t)$ as a source property, we can investigate the consequences of $T(t)$ as a continuous function of time. On the other hand, if it is known or hypothesized that a segment sequence $s(t)$ represents information of a single signal component or sound event, it is possible to use the TNC to study the development of information represented by the running autocorrelation under the segment sequence $s(t)$.

2.6 Estimation of Ridges

The instantaneous local frequency information as represented by the TNC forms the basis for the estimation of pitch-contours in unknown noisy circumstances. Computationally, the TNC is extremely inefficient since it is of the order (# segments) times (# samples per second) times (# periods). For 100 segments, a sample frequency of 20 kHz and a maximum period of 25 ms (500 different values) this corresponds to $10^3 \times (2 \text{ multiplications} + 1 \text{ addition})$ per second. Although it is possible to increase the efficiency of the computation considerably, a more efficient approach is required.

8. According to theorem 1.21, efficient FFT-based approximations of the TNC (Slaney 1993) are not allowed for signals that are not (yet) recognized.
Fortunately, it is easy to determine regions in the cochleogram that are likely to provide prominent information about a single signal component (e.g., a harmonic). As discussed in the context of figure 2.4, each signal contribution tries to entrain a region of the basilar membrane. A periodic signal component that succeeds to dominate a region of the BM will lead to a peak at the BM position corresponding to the frequency of the signal component. This means that peaks usually correspond to a single strong signal component. Relatively weak signal contributions like the 10th to 12th harmonic in figure 2.4 are almost completely masked by stronger contributions and do not show up as separate peaks. When the search space is limited to peaks in the cochleogram, one efficiently selects positions where information of entraining signal components can be estimated!

To reduce the number of spurious peaks, ridges can be formed by combining peaks through time. All peak-positions that cannot be classified as members of ridges equal to or longer than 20 ms (4 frames of 5 ms) are discarded. This leads to figure 2.7, which allows a comparison between ridge estimates in noise with estimates in the clean situation. The left hand panel gives the ridges as estimated in 0 dB cocktail party noise superimposed on the cochleogram of the clean target signal /NUL/. The right hand panel shows the complementary information: the noisy cochleogram with the ridges as estimated in a clean signal. The ridges estimated in the noisy signal often coincide with the most prominent peaks of the clean target. Since these ridges are estimated from a noisy signal they include positions where signal components of the background dominate.

As can be seen in the right hand panel, the cocktail party background consists mainly of the intensity peaks of the speech of other speakers. Since these intensity peaks last shorter than the whole sound event, the requirement of ridges of at least 20 ms removes an important fraction of the background peaks. For backgrounds consisting of a lot of uncorrelated sources, or backgrounds containing aperiodic noises this is often the case. This requirement helps to alleviate the problems associated with the signal-in-noise paradox by applying knowledge of the target signal: in this case the fact that speech usually consists of contributions of at least 20 ms. This constraint

---

9. Note that although the word frame is used to denote a signal-block of 5 ms, quasi-stationarity by frame-blocking is not applied. The basic signal representation conserves continuity, but some features, like peak positions and the local instantaneous frequency are sampled each 5 ms.
reduces the search space by discarding nonspeech contributions much more often than the target speech.

### 2.7 Local Instantaneous Frequency Contours

The next step towards a robust pitch estimation algorithm is the estimation of *Local Instantaneous Frequency* (LIF) contours. We now have a set of continuous ridges \( \{s_i(t)\} \) and since the TNC is continuous in time \( t \) and place \( s \), it is possible to compute the running autocorrelation along the ridge \( s(t) \) as:

\[
 r_{s(t),T}(t) = r_{s(t),T}(t - \Delta t) e^{-\frac{\Delta t}{T}} + x_{s(t)}(t) x_{s(t)}(t + T) \quad T = 0, \ldots, T_{\text{max}} (2.10)
\]

As the peak position changes smoothly, so does its associated autocorrelation. Note the symmetry with the tuned autocorrelation of equation 2.8. That equation represented a set of functions over all segments \( s \) with period contour \( T(t) \) as a function of time, while equation 2.10 is a set of functions over
all $T$ with the segment sequence $s(t)$ as a function of time. The TAC describes vertical cross-sections of the TNC and the running autocorrelation a horizontal cross-section.

Figure 2.8 gives examples of a few autocorrelations estimated a time $t=250$ ms (see figure 2.7) from the noisy /NUL/. The lower panel shows the running autocorrelation of the ridges 2, 4, 6, 7, and 8 (numbering starting from lowest ridge) at the position of the bar in figure 2.7. Since they all agree on a periodicity of approximately 4.10 ms, these ridges might reflect harmonics of a single source. The upper panel shows the autocorrelation of ridges 1, 3, and 5 that do not agree with this periodicity.

The local instantaneous frequency is related to the inverse of the period corresponding to the position of the first peak in the autocorrelation. These values are computed and depicted in figure 2.9 for two conditions: the blue
stars are the values of the local instantaneous frequencies as estimated from the clean /NUL/. The red circles are estimated from the noisy /NUL/. Note that most frequency contributions in the clean signal remain clearly present in the noisy environment. A closer examination shows that perturbations are often less than 2 percent. This shows that the ridges form a very reliable source of information for the estimation of the frequency development of signal components.

Since the basilar membrane model (as used in this work) has only 100 segments, which span more than 2 orders of magnitude in frequency, the estimate of the instantaneous local frequency based on the position of the ridge(s) cannot be very accurate. The estimation of the instantaneous local frequency based on the running autocorrelation is accurate to within less than a percent for frequencies far below the Nyquist-frequency of the input (here 10 kHz). This is surprising because the running autocorrelation has an

---

10. Actually a somewhat more complicated and more accurate regression method is used that takes the rest of the running autocorrelation into account as well. This is discussed in section 4.5
associated exponential window with a time constant of 10 ms. One could argue that the resulting time-frequency trade-off ought to limit frequency resolution to approximately 100 Hz. However, one can easily make a distinction between quasistationary signal components of 100 and 101 Hz (see section 4.5). The time-frequency trade-off is determined by the BM model (and not by the exponential window) and the source that produced the signal. Because the BM model has an infinite window of integration it can produce frequency estimations that are only limited by the development of the target signal. This is another illustration of the benefits of the avoidance of frame-based approaches.

2.8 Design Choices and Overview of Implemented System

In the last part of this chapter the target signal is extended to /NULEN TWEE DRIE/(English: /ZERO ONE TWO THREE/).\textsuperscript{11} Cocktail-party (or babble) noise is added so that a signal-to-noise ratio of 0 dB results. The associated cochleogram and the estimated ridges are depicted in the upper panel of figure 2.10. The lower panel shows the frequency contributions along the ridges. (Note the difference in scaling of the frequency axes.) An audio presentation of this sentence can be recognized without much difficulty by speakers of the Dutch language. Yet at -3 dB, naive listeners are unable to recognize the noisy /DRIE/ when it is presented without the preceding digits. When these are added to the signal, the /DRIE/ is usually perceived clearly. This suggests the importance of linguistic context at this signal-to-noise ratio. At -6 dB, listeners are often unable to detect the sentence at first presentation. This thesis uses the target sentence of figure 2.10 because these informal experiments suggested that an SNR of 0 dB (in the case of babble-noise) is just above the threshold that prevents speech to be correctly processed without additional information from linguistic context, binaural hearing, visual cues etc.

Figure 2.10 shows the information available for the estimation of fundamental period contour hypotheses. The use of ridges, ensuring an efficient reduction of the search space, and the availability of mostly accurate estimations of the local instantaneous frequency must facilitate this process. Yet the robust

\textsuperscript{11} Will be made available at http://www.bcn.rug/andringa/thesis
estimation of the fundamental period contour, or its corresponding pitch-contour, is surprisingly difficult: often there is no unique solution. This is a direct consequence of the limits of the measurement process (see page 31). The optimal solution is to work with a set of hypotheses, estimated to ensure that it contains, as often a possible, enough information to allow a correct recognition result. Then a full search through this set is performed to find the hypothesis combination that explains the data best. Unfortunately, speech recognition systems based on this approach are not yet available (but section 7.2 proposes such a system).

At this stage an important and suboptimal design decision is made. In order to use a standard HMM-based speech recognition system for testing the benefits of the selection process, a single sequence of the target signal must be produced.

This entails that the system needs to find the best fundamental period contours while limiting itself to the information in the signal. Since no linguistic knowledge can be applied, the system will be unable to detect pitch estimation errors and nonspeech sounds.\(^\text{12}\)

Figure 2.10. The upper panel shows the cochleogram of /NUL EEN TWEE DRIE/ with added cocktail party noise at a signal-to-noise ratio of 0 dB. The lower panel shows the frequency contours associated with the ridges in the upper panel. Note the difference in scaling of the frequency axis.
Finding a unique and correct pitch-contour becomes progressively more difficult and eventually impossible when the signal-to-noise ratio decreases. Furthermore, problems arise in situations with multiple speakers and/or music.

To facilitate implementation even more some other, suboptimal, design decisions are made. One choice is to develop a proof-of-concept with parameter values that have not been optimized in any formal way. Another important design choice is to use the whole signal instead of the most recent information. This introduces a delay that effectively entails that the system can only be used in a multistage mode. The first stage computes the basilar membrane response, the ridges and the instantaneous frequency contours of the whole input signal. The second stage computes the fundamental period contour, performs selection with the TAC and produces a parametrization suitable for the last stage that computes the recognition result. This leads to the following system overview:

This is, of course, far from the ultimate goal: a real-time system that can correctly process sounds of which a priori knowledge might or might not be available. Nevertheless, a system like this will be useful for applications that have to deal with variable and uncontrollable signals, e.g., telephone-based information systems.

The rest of this chapter is not aimed at building such a system, but aims to develop a signal processing approach that is suitable for variable acoustic environments. A suitable test of this system differs slightly from the evaluation of the process described in the system overview. It requires an evaluation of the behavior of the system when a correct pitch-contour is available since this gives an indication of the quality of the preprocessing. It is, conform conclusions 1.8 and 1.13, the task of later recognition stages to identify the correct pitch hypotheses on the basis that these lead to an

---

12. Note that this is a consequence of the separation of selection and recognition in two separate processes. The more integrative approach of section 7.2 avoids this problem.
acceptable recognition result with a meaningful interpretation. Although pitch estimation in noise might become fairly reliable, it will never be possible to guarantee a correct choice between two or more well-supported pitch hypotheses without knowing which choice leads to the most meaningful interpretation.

2.9 Fundamental Period Contour Estimation

The development of a reliable and robust pitch estimation technique is difficult. The main reason for this is that it is generally impossible to determine which signal contributions or signal properties belong to a certain source prior to recognizing the sources. This is a direct consequence of the inability to determine whether a signal is speech or not without being able to recognize the signal that was discussed in section 1.1. Yet although this problem is generally insoluble, some features, like smoothly developing harmonics, can be used to search for evidence of speech. Although these features are shared by other sounds than speech, they can be used as long as the user ensures that other types of sounds do not occur (which is, unfortunately, not what one requires for a system that can deal with unknown situations).

Two fundamental period contour estimation techniques were developed: one for clean signals, and one for noisy signals. The first uses the fact that the target is not contaminated with noise. It is reliable, but sensitive to added noise. It is based upon the property that all harmonics of a periodic source show a common periodicity, as is shown in the lower panel of figure 2.8. This technique is used to determine the quality of the TAC selection technique in section 2.13. A more complete description of this method will be postponed to section 5.2. The second period contour estimation technique is, conform the limitations on page 63, developed for a larger class of signals. This technique is not developed to be flawless (which is impossible), but intended as a proof-of-concept: it has to come up with an approximately correct pitch-contour most of the time. An overview of this technique will be given in this section. A more detailed description can be found in section 5.1 on page 140.

A first choice is whether to work in the period domain or the frequency domain. This is an arbitrary choice because both domains are equivalent, but since the final result of the technique is a period contour, two domain changes can be avoided. The available information is depicted in the upper panel of
For visual clarity only periods smaller than 5 ms (i.e., frequencies higher than 200 Hz) are depicted. Since this representation is based on a situation with a signal-to-noise ratio of 0 dB it shows a lot of spurious contributions that must be eliminated somehow.

To attack this problem, a set of heuristics are required. Inspection of noisy speech signals showed that the most energetic, the longest (longer than 50 ms) and the smoothest ridges often reflect the most reliable sources of information. These heuristics are used to select a number of points of the upper panel of Figure 2.11. The selected points form ridges of segments that on average belong to contours longer than 75 ms, or to contours longer than 50 ms that, on average, contain the (two) most energetic segments per time step. The selected ridges are substituted by a smooth approximation that is just as long. These ridges are depicted in the lower panel.

Figure 2.11. The upper panel shows raw periodicity information. This is the inverse of the information as depicted in the lower panel of Figure 2.10. The lower panel shows the result of an algorithm that selects long contours from the upper representation and smooths the final result. These form the seeds for the next stages in the estimation process of the best fundamental period contours.
The smooth ridges $p(t)$ might, or might not, stem from harmonics of the target speech. If the harmonic number $n$ would be known the fundamental period $p_0(t)$ would be known, since:

$$p_0(t) = p(t)n \quad \text{or} \quad f_0(t) = \frac{f(t)}{n} \quad (2.11)$$

As a further limitation, valid fundamental period values are limited to values between 2.5 ms (400 Hz) and 13.3 ms (75 Hz), a range that spans most speakers (Furui 1989). For example an instantaneous period $p=6$ ms can be the result of the second harmonic of a fundamental period $p_0=12$ ms, or the first harmonic of $p_0=6$ ms. A period $p'=2$ ms can represent any harmonic number in the range of 2 to 6. This corresponds to any $p_0$ in the set {4, 6, 8, 10, 12} ms. If $p$ and $p'$ stem from the same source, they share the same fundamental period $p_0$. In this case either 6 or 12 ms.

This property is used for the contours as depicted in the lower panel of figure 2.11. The upper panel of figure 2.12 shows all fundamental period contour combinations consistent with the smoothed contours of figure 2.11. Note that some hypotheses overlap. The lower panel shows a selection of the contours of the upper panels based on length and smoothness. The fine structure in some of the lines of the upper panel is an artefact of the plot program.
hypotheses consistent with the smoothed contours of figure 2.11. Some of the fundamental period contour hypotheses overlap or extend each other smoothly. This is a strong indication that the period contours stem from the same source: the probability that uncorrelated period contours form a consistent whole is small, but not zero! The lower panel depicts a selection of the upper panel based on three main criteria: the contours must have a certain minimal length (50 ms), they must be sufficiently smooth and in case of multiple concurrent contours only the longest contours are selected. This results in a strong reduction and it often results in a set that includes a more or less correct pitch-contour candidate.

The final step compares the remaining concurrent candidates with the original local periodicity information, depicted in figure 2.11, to determine which candidate explains most of the period values and, to prevent octave errors, has a reasonable ratio of odd and even harmonics. The candidates that meet these demands best forms the final output of the algorithm. Figure 2.13 shows a comparison between pitch-contours estimated from signals with different signal-to-noise ratios of babble noise. Apart from some differences during on-
and offset, the algorithm is able to find the correct contours for SNR's of -3 dB and better. When the algorithm produces a correct contour, the match is usually well within 1% of the actual value. This is not surprising since the most prominent harmonics of the target sounds are still quite able to dominate locally in these conditions. The algorithm identifies these regions and uses periodicity information to find the pitch-contour that combines as many of these regions as possible. Because the periodicity information is still almost unimpaired, the pitch-contour must be of similar quality as estimated in clean conditions. During onset and offset the energy of the target is generally lower which may lead to an unfavorable local signal-to-noise ratio. This makes it more difficult to determine the period contour unambiguously.

Since the pitch-contour estimation technique looks for long, smooth and well supported fundamental frequency contours, it finds all combinations of evidence that can be supported. These may, or may not, correspond to the true pitch-contours. Below an SNR of approximately -5 dB babble, speech, or car factory noise the algorithm is unlikely to function reliably because it becomes impossible to form a set of hypotheses that include the correct period contours. As said before: this version of the pitch estimator is intended as a proof-of-concept. It can be improved and eventually it ought to be replaced by a version that produces multiple hypotheses that are evaluated by higher levels of processing.

2.10 Selection of Periodic Signal Contributions

The next step is the actual assignment of information to an auditory event-like representation. The lower panel of figure 2.14 shows typical examples of TAC-based auditory events. It is surprising to see what the application of a single constraint, the period contour, can do with the noisy signal in the upper-panel. On the low-frequency side, the TAC cochleogram selects the first harmonics reliably, on the high-frequency side it selects large areas of the time-place plane. On the low-frequency side the selected regions are dominated by a single harmonic. On the high-frequency side the regions are dominated by formants: complexes of harmonics that agree on a common fundamental period.
It is now evident that the TAC contributes to the solution of the signal-in-noise-paradox (definition 1.5). Furthermore, the power of the TAC-approach illustrates the meaning and feasibility of the theoretical solution of the signal-in-noise-paradox as formulated in conclusion 1.12:

The signal in noise paradox can be solved by grouping continuously developing acoustic information of a single source into auditory events. A search through the set of auditory event combinations might produce a number of acceptable recognition results.

Although the selection as depicted in figure 2.14 was based on correct period contours, it cannot be guaranteed that the selection is correct: one of the background speakers might be the source of one of the period contours. Section 4.7 addresses this problem in more detail. Further processing, using knowledge of speaker characteristics and all aspects of language, must solve this problem. Fortunately, the information represented by an auditory event, based on a correct period contour estimated in rather noisy situations,
comprises accurate information about the relative importance of individual harmonics and formants. This is enough to reduce the number of interpretations to a few hypotheses.

Although the TAC-approach cannot assign non-periodic information to auditory events, it can help in determining the position of likely candidates of aperiodic auditory events that might be assigned to the same stream. In normal speech the position of aperiodic signal components is strongly correlated to the periodic components. In most cases, these contributions end just before or during the onset, and start at or after the offset of a periodic contribution. In the case of the /T/ of /TWEE/ (/TWO/), starting at t=1000 ms and most noticeable in the segment range from 90 to 100 in the upper panel of figure 2.14, some form of template matching in combination with the Characteristic Period Correlation (section 4.3) may suffice to detect and characterize likely candidates of aperiodic contributions.

2.11 Resynthesis of Target Signal

Because the TAC forms a reliable basis for the assignment of information to auditory events, one might ask whether it could be used to split a combination of sounds into the constituting sound events. This is, within the limits of the discussion in section 4.7, possible and is in fact a straightforward procedure. All quasiperiodic signal contributions that dominate a certain region in the time-place plane of the TAC cochleogram represent basilar membrane oscillations. Since the basilar membrane model is implemented as an impulse response-based finite impulse response (FIR) filter, it is possible to inverse the filtering by reversing the impulse response in time and compensating for the frequency-effects caused by the double use of the basilar membrane filter (Slaney 1994).

A full inversion results in the original mixture of signals. But if inverse filtering is based on the regions of the time-place plane that are dominated by the target source, the output is, ideally, exclusively based on information from the target. If all positive values of the TAC cochleogram of figure 2.14 are used as a mask the result sounds unpleasant. This is mainly due to on- and offset effects of filtering and to the contribution of incidental correlations. To reduce the latter effect, TAC-values smaller than a certain fraction, e.g., 0.25, of the local instantaneous energy are discarded. The remaining TAC-values are depicted in the upper panel of figure 2.15. To reduce the effects of on- and
offsets the mask is tailored to consist of long (here minimal 50 ms) continuous
contributions of single segments: small holes (here 20 ms) in the positive
values of the TAC-traces are filled up and isolated positive points are
discarded. Finally, the mask is provided with smooth 10 ms wide on- and
offsets. This leads to the mask as depicted in the lower panel of figure 2.15.

To improve the sound quality, the background is not completely discarded,
but reduced with an adjustable factor: in this case a factor of 100 in amplitude
(40 dB in terms of energy). By not completely discarding the background,
unnatural ‘deep’ silences are reduced and some evidence of aperiodic
contributions, like the /T/ of /TWEE/, remains in the signal; this facilitates
perception. When the resulting resynthesized sound is again presented to the
basilar membrane model, the cochleogram of the resynthesized sound can be
computed. This is presented in the middle panel of figure 2.16. The upper
panel shows the cochleogram of the original signal. This signal formed the
only source of information: no a priori information was used, nor necessary.
The lower panel shows the clean reference. Apart from the second formant

Figure 2.15. The upper panel shows the regions where the TAC-value is larger
than 25% of the local energy. The lower panel shows the mask that is used as a
template for resynthesis. The mask is based on the positive values in the upper
panel, but tailored to consist of segment contributions of a certain minimum
length. The light points at the end of the lines signify the tapering to further reduce
on- and offset effects.
structure of the last word, which is masked completely, all important periodic contributions are represented faithfully. Note that the resynthesized cochleogram is more “fuzzy”, this is due to spurious contributions of the background. One way to avoid this is to measure and smooth all individual signal components and add these together in a true speech synthesis process.

Although the intelligibility is high, the perceptual quality of very noisy (SNR 0 dB) resynthesized signals is not very good. This is because quite a lot of the background, visible as the fuzziness in the middle panel of figure 2.16, is still present in the selection. The original signal provided our auditory system with enough cues to separate the signal components and assign individual

Figure 2.16. The cochleogram of the resynthesized signal, shown in the middle panel, reflects most features of the original signal in the lower panel. The representation of the middle panel was entirely based on the noisy signal shown in the upper panel. The fuzziness in the reconstructed signal is due to spurious signal contributions of the cocktail party background. The /T/ starting around $t=1.0$ s is absent in the reconstruction since it is unvoiced.
signal components, on the basis of perceptual coherence, to the correct source. In
the resynthesized signal the majority of these cues are removed, which makes it more dificult to assign spurious contributions to other percepts. Consequently, the perceptual quality of the signal is low. For true speech enhancement, further optimizations of the resynthesis process are required. An improvement is to estimate relevant signal properties, such as the pitch-contour, the development of the first harmonics and the position and development of the second and third formants, and then use this information as input for a speech synthesis system. Techniques like this may lead to high quality coding at very low bit-rates. In its current form, resynthesis is mainly suitable for visual purposes. Mask forming and resynthesis will be improved in sections 6.1 and 6.2.

2.12 Parameterization of Selections

This section and the following show that the information in the TAC-selections can form a suitable basis for robust HMM-based ASR systems. The resynthesized sound, as computed in the previous section, can be used as input for a standard off-the-shelf speech recognition system. Unfortunately, this resynthesis does not contain any information about unvoiced (aperiodic) phonemes and it is dificult to predict how a standard, pretrained, ASR system will respond to these signals. A practical problem is that the resynthesis procedure requires an extra pass through the basilar membrane model. Since the basilar membrane model forms the computational bottle-neck of the system, its application ought to be minimized. Instead, the input of the recognition system can be based on the TAC-cochleogram.

A suitable input for an ASR system is a description of the temporal development of the spectral envelope of the target speech (see figure 1.1) while suppressing pitch effects. As the upper right-hand panel of figure 2.5 shows, the TAC-cochleogram of the voiced parts of a clean signal resembles the standard cochleogram closely. The TAC-cochleograms in the lower panel of figure 2.14 resemble the clean cochleogram better when the negative values are replaced by suitable positive values. The TAC-approach allows an (usually) accurate estimate of the energy and frequency development of the

---

13. The results of more elaborate and convincing recognition experiments on the standardized Aurora test for robust speech recognition are available at: http://www.huq.nl.
first few harmonics. Because voiced speech mainly consists of a superposition of harmonics and because the TAC-approach is linear, it is possible to recreate the lower part of the cochleogram by adding contributions of the ideal responses of individual harmonics. This superposition does not contain any negative parts. This procedure is introduced in section 3.4 and described in detail in section 3.6.

The information in the upper part of the cochleogram, where individual harmonics cannot be resolved, is based on masks as computed in the previous section. The information under the mask is kept unchanged. Outside the borders of the mask, vertical “tails” are added to reflect masking upward and downward in frequency. As a last measure a (horizontal) “tail”, as an approximation of the ringing-out effect of the basilar membrane and the leaky integration process, is included.

This results in synthetic correlograms as depicted in figure 2.17. The upper panel shows the “reconstruction” based on the TAC of the clean signal. A comparison with the lower panel of figure 2.16 shows that the main components of both figures are very similar. This indicates the validity of the reconstruction method. The lower panel of figure 2.17 shows the reconstruction based on the TAC as estimated from noisy data. Since part of this signal is masked and some spurious contributions of the background are added, the match is not perfect, but the main features of both figures are similar (under a visual inspection). Prior to a correct recognition it is unknown which of the contributions are spurious, and consequently it is impossible to make a perfect selection. This problem can only be alleviated by the application of more and more characteristic constraints; which is not possible within traditional ASR-systems.

An HMM-based ASR system requires an estimation of the spectral envelope of the target speech without pitch effects. The representation as depicted in figure 2.17 is not very suitable since the first harmonics are the most energetic components. Although these carry formant information, the detailed realization of the first formant depends strongly on pitch. To reduce the effect of irrelevant pitch differences and to stress the second and third formant, a very simple and quite arbitrary trick is used: the values of the compressed cochleogram are multiplied by a segment dependent factor. This factor is 1 for the first segment and maximal, e.g., 5, for the last segment. The multiplication factor of intermediate segments is a linear interpolation between the two extremes. This is an operation with a similar effect as pre-emphasis, a form of high-pass filtering that is usually applied within the standard methodology of
ASR, and results in a speech spectrum where all frequencies contribute on average with the same amount of energy.

As a final step, the envelope of the cochleogram must be coded as efficiently as possible. To produce a set of parameters similar to MFCCs, a cosine transform of the weighted cochleogram is performed. The result is a variant of a standard cepstrum. Typically the first 8 to 14 values of the cepstrum, representing low spatial frequencies are kept, the rest is discarded. Finally, the

14. MFCC: Mel Frequency scaled Cepstral Coefficient. MFCCs are a well-known FFT-based representation involving an approximation of the place-frequency relation of the basilar membrane and a logarithmic compression. The envelope structure is derived with a cosine transform. The resulting representation is termed a cepstrum. The values that represent the contributions of each cosine-base vector are called cepstral coefficients (Rabiner 1993, Gold 2000).
time-step between successive frames is increased from 5 ms to 10 ms by averaging successive values. This makes the frame step a standard value and speeds up processing. These values are stored and used as input for the speech recognition system.

The stored parameters are not very informative, but they can be transformed back to a cochleogram-like representation by applying the inverse cosine transform. The result is shown in figure 2.18 which reflects the information available to the speech recognition system. The upper panel is based on the original clean signal. The energy contributions per segment are enhanced by values between 1 and 5, the spectral envelope is coded with 12 cepstral coefficients. Compared to the lower panel of figure 2.16, the high-frequency segments are much more prominent, the first harmonics are less prominent, and the formant features are broader. The lower panel is based on the reconstructed TAC-cochleogram of figure 2.17 and has a good general agreement with the ideal cochleogram, but it is noisy due to masking and
spurious background contributions. These two representations form the basis for the recognition experiments of the next section.

### 2.13 Recognition Experiment

This chapter set out to prove that the basic framework presented in chapter 1 is useful for the improvement of a standard speech recognition system. So far, a selection technique for quasiperiodic sound events has been developed that appeared robust under visual inspection. The final proof is whether TAC-based features are robust to added noise. Preliminary experiments with a standard ASR system gave an indication of the robustness of the TAC-approach. This section provides, conform the design decisions of section 2.8, an overview of recognition experiments with period contours estimated from noiseless signals. The way these period contours are estimated is outlined in section 5.2.

For the ASR experiments the Hidden Markov ToolKit (HTK, version 2.2, Young 1999), a well known state-of-the-art ASR development system was used. The ASR test is limited to the 5414 connected digits strings of the female training data of the TI-DIGITS data-base (Linguistic Data Consortium 1993). Both HTK and the TI-DIGITS task are well known and are often used for benchmarking and comparison. The recognition system uses 12 whole word models: /ONE/ to /NINE/, /ZERO/, /OH/ and /SILENCE/ and recognizes digits strings like: /SILENCE ONE FOUR SEVEN TWO ONE SILENCE/. A potential problem is that the TAC cannot select unvoiced phones like /TH/, /Y/, /S/, /K/, /F/ and it will select only the periodic parts of phonemes like /Z/ and /V/. The most difficult digit is the /SIX/ since it consists of two unvoiced parts while the voiced part of the /I/ is often short (even less than 50 ms) and of relatively low intensity. This makes it hard to detect the pitch reliably.

Two speech recognition systems were built: one with standard MFCC coefficients and one based on TAC-selection with the parameterization of section 2.12. Both were trained on clean data and tested on the trainings data contaminated with added noise. The test used babble noise, car factory noise, speech noise and white noise at -5, 0, 5, 10, 15 dB SNR of the NOISEX database (Varga and Steeneken, 1992). For comparison, recognition tests on the clean signals were performed as well.
Both systems are based on continuous density HMMs with diagonal covariance matrices and 10 states with self loops but no skip transitions. The input consisted of 12 cepstral coefficients, their temporal derivative (or delta's), the energy and the first order derivative of the energy. Training was based on a flat-start (all states filled with the global average of the training data). Baum-Welch reestimation was applied to bootstrap the models. The models were refined by adding two mixtures at the time followed by two Baum-Welch reestimation passes. Eventually the models consisted of 15 mixtures. The benefit of the last 6 mixtures is minimal, but somewhat more important for the selection-based system.

The TAC-based system is expected to be more robust than the MFCC-based systems. But when tested on clean input, the MFCC-based system is expected to outperform the TAC-based system, since the latter ignores the unvoiced phonemes and involves more processing steps that can all introduce an error. Yet at some point along the signal-to-noise ratio continuum these disadvantages ought to be overcome by the benefits of selection.

The results in figure 2.19 confirm these expectations. For the matched condition, i.e., testing on clean speech, the performance of the MFCC system is 99.3% word correct, but performance drops rapidly for decreasing SNR. The TAC-based system reaches only 94.3% for the matched test, but performance is much less sensitive to added noise. The rather poor performance can be attributed to several causes.

- Lack of optimization. These results are a first attempt, and no optimization whatsoever is performed. Since almost every aspect of the system can be improved this is probably the main cause.
- Absence of unvoiced signal components. This might be an important factor since the a recognition system trained on a cepstral representation of the cochleogram (including all voiced peaks) reaches a performance of 99.1%.

15. Note that this test does not aim to establish the quality of a recognition system but measure the robustness to noise of the parameterization. This kind of task requires that the training and testing database are equal. Experiments show that an unmatched test leads to a minimal degradation, which shows that the system has not been overtrained (i.e., its parameters do not reflect the peculiarities of individual samples in the training set).

16. See Bourlard 1995 (Towards increasing speech recognition error rates) for an other form of justification of these scores.
Figure 2.19. The results of the recognition tests confirm that the TAC-selections lead to a robust representation of speech. The recognition performance of the TAC-selections in clean situations is below the results of the matched MFCC-test (trained and tested on clean speech). But a 10% average performance loss compared to the matched condition occurs for the MFCCs around +20 dB SNR, while for the TAC-selections the 10% performance loss occurs at an SNR of +4 dB.

- Missing pitch-contours. The selections are based upon offline computed pitch-contours. This algorithm is not flawless, which means that some contours, especially for the /SIX/ and the /EIGHT/, were not estimated, while other contours might be a fraction to short or interrupted. This might explain up to 10% of the errors (or 0.6% in terms of percentage correct). See section 5.2 and figure 5.4 for the basis of this argument.

- Pitch effects in the parameterization. As figure 2.18 shows, the first and second harmonics are still noticeable in the inverse cepstral representations, this means that a considerable fraction of input information represents pitch information instead of envelope information. Since pitch information does not contribute to word identity estimation (in English), the HMM models will not represent word identity information efficiently.

- Independence assumption of HMMs. A cochleogram-based MFCC representation of speech is smoother and therefore more predictable than standard MFCCs. This entails that the TAC-based parameterization is more dependent (in a statistical sense) than MFCCs. Since HMM-based
Recognition Experiment

systems assume statistical independence of consecutive frames (see the footnote on page 19), the TAC-based parameterization violates this basic assumption of HMMs more than standard MFCCs do. This effect is difficult to quantify and will be the subject of further research.

Although the TAC-based parameterization might be far from optimal, it is very robust to different types and levels of added noise. There is little performance gain for 5 dB and better. This suggests that the parameterization changes little in this range. This supports the visual evidence of the previous sections. Furthermore, the selection is only weakly dependent on the type of added noise, this proves that, as claimed in section 2.4, the TAC selects the target much more efficiently than the noise. Finally, the recognition performance for very adverse signal-to-noise ratios is still high. In an SNR of -5 dB more than 50% of the digits are recognized correctly (chance level is below 10%). This result is positively influenced by the availability of the correct period contours. Nevertheless, an SNR of -5 dB is a very challenging situation that requires full attention of a human listener.

One might take the SNR corresponding to 10% performance loss compared to the matched condition as a measure of robustness. This level is depicted in figure 2.19 as the dashed line. For the MFCCs, the point of 10% performance loss is around a SNR of 20 dB, the exact point cannot be estimated due to the limited number of tested noise levels. For the TAC parameterization, the 10% performance-loss-point lies, on average, at +4 dB.

It can be concluded that the application of the tuned autocorrelation may lead to a considerable increase in the noise robustness of a standard speech recognition system. The performance in clean situations can be improved by an optimized recognition strategy and by the inclusion of aperiodic speech components.