Continuity preserving signal processing

Andringa, Tjeerd

CHAPTER 7 Overview and Discussion

This chapter discusses CPSP in the context of the theoretical framework of chapter 1. Section 7.1 starts with an overview of the main representations, their properties and some possible extensions. Section 7.2 proposes a method to apply these techniques to speech recognition in much more variable environments than is possible with current HMM-based technology. Finally, section 7.3 argues that the use of conjectures 1.1 and 1.2, on which this work is founded, leads to a system with properties consistent with human performance, and it draws some conclusions about the advantages and application domains of CPSP.

7.1 Overview of CPSP

The main goal of this work is, conform conjecture 1.1, the formulation of a signal processing framework that allows recognition systems to function as often as possible in varying and uncontrollable acoustic environments. Following conclusion 1.14, the approach has been to start from the weakest possible prior assumptions. For sounds, the weakest possible prior assumption is that sounds consists of signal components that each show an onset, an optional *continuous development* and an offset (definitions 1.9 and

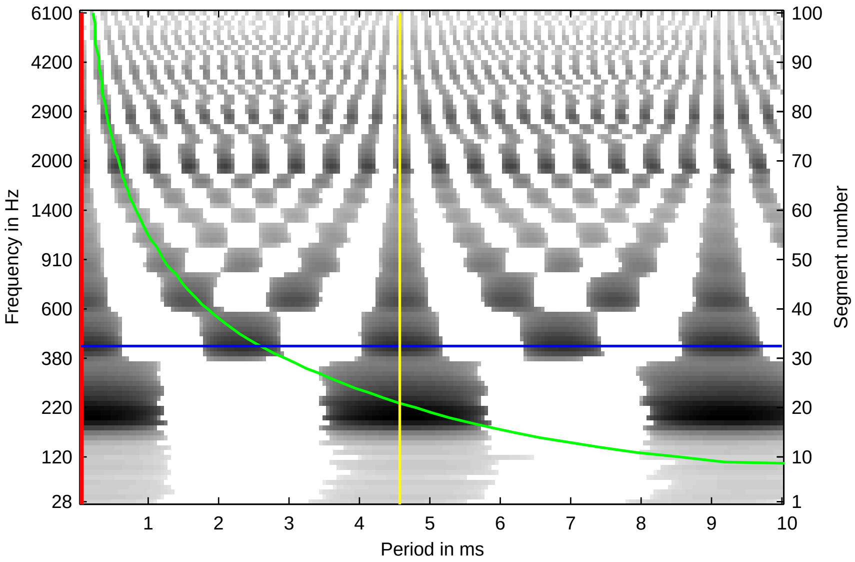

Figure 7.1. Overview of the relations between the TNC and four special subsets. The red line at T=0 corresponds to the cochleogram, the yellow vertical line at T=46 ms marks the tuned autocorrelation, the green line corresponds to the characteristic period correlation (according to definition 4.7) and the horizontal blue line at segment 32 reflects the running autocorrelation along a ridge. The TNC structure in the background is an example of the periodic TNC.

1.20). Consequently, the decision was made to preserve continuity as long as possible and to postpone the application of quasi-stationarity to the moment it can be justified (conclusion 1.21). This led to the use of a transmission line model of the basilar membrane, the formulation of the *cochleogram* and its generalization that includes periodicity: the *Time Normalized Correlogram*. The TNC always reflects a superposition of two qualitatively different stable patterns: one associated with the *aperiodic excitation* of the corresponding BM region, the other associated with a *periodic excitation*. This allows the analysis of signals in terms of periodic and aperiodic components, for example to separate on- and offset transients from the steady state behavior (section 4.4).

Some subsets of the TNC provide special information about the state of the BM. Figure 7.1 depicts examples of these subsets as lines superimposed on a typical periodic TNC. The red line at T=0 corresponds to the cochleogram and reflects the energy of the basilar membrane as a function of time and place. The yellow vertical line at T=46 ms marks the *tuned autocorrelation* and is related to the periodic excitation of the TNC. When a correct period contour is known (or can be estimated) the TAC has values similar to the local energy for all target source harmonics that dominate BM segments.

The green line connects the characteristic frequency of each segment (vertical axis) to the characteristic period (horizontal axis) and marks the *characteristic period correlation*. The CPC is derived from the aperiodic excitation pattern of the TNC (actually its impulse response). Basilar membrane segments that oscillate with the local characteristic period indicate the presence of a strong signal component. The ratio between the CPC and the local energy is a measure of *local dominance* (section 4.3). For dominant *aperiodic* signal components the CPC was defined to be close to unity, for dominating *periodic* signal components the CPC can be greater than unity.

Periodic components of sufficient duration and sufficient (relative) intensity entrain BM regions and lead to *ridges* (section 3.5). The horizontal blue line at segment 32 reflects the running autocorrelation along a ridge. This running autocorrelation allows a maximally accurate estimation of the development of the *Local Instantaneous Frequency* (LIF) of the individual dominating periodic components (section 4.5).

Table 7.1 provides an overview of these representations and their main properties. All representations are variants of a general correlation function $r(s,t,t_2)$ that depends on BM position s, running time t and period t_2:

$$r(s, t, t_2) = L\{x_s(t)x_s(t + t_2)\} \tag{7.1}$$

with L a lowpass operator. And:

$$x_s(t) = b_s(y(t)) \tag{7.2}$$

Where $x_s(t)$ the output of a cochlea model $b_s(t)$ at time t. These representations allow the estimation of *auditory elements* A_i according to set inclusion criteria of the general form:

$$A_i = \{(s, t) \mid f_i(r(s, t, t_2)) > \theta_i\} \tag{7.3}$$

An auditory element of type A_i is set of spatio-temporal points (s,t) that exceed a threshold θ_i according to a decision function $f_i(r)$ that may be based on combinations of different choices of $r(s,t,t_2)$. Figure 7.2 provides a system overview of a *Continuity Preserving Signal Processing* system, with in the boxes the requirements to be satisfied and below the boxes some optional featur es.

The input $y(t)$ of a CPSP system is not restricted in any way. The basilar membrane function $b(y)$ is characterized by a continuous and invertible place

General form: $r(s, t, t_2) = L\{x_s(t)x_s(t + t_2)\}$ $x_s(t + t_2) = b_{s, t + t_2}(y(t))$	$r(s,t,t_2)$: correlation at position s, running time t and with period t_2. $y(t)$: input sound $b_{s, t + t_2}(y)$: BM output of segment s at time t+t_2. $L_s\{\ \}$: lowpass filter, slow compared to $x_s(t)$.
Time Normalized Correlogram (TNC): **$r(s,t,t_2)=L\{ x_s(t)\ x_s(t+t_2) \}$** Sections 2.5, 4.1 and 4.2.	Time-of-onset normalization. Synchronization of response with period input. Structure of impulse response of TNC is time-invariant. TNC regions shows a combination of periodic responses and the impulse response TNC.
Cochleogram: **$r(s,t, 0)=L\{ x_s(t)\ x_s(t) \}$** Sections 2.3, and 3.2 to 3.4.	Measure of energy as a function of time and place. Approximation of periodic cochleogram cross-section: R(t)={r(s,0,t) $\forall s$ }: $R(t) = \sum w_n(t)R_n(t)$ (eq. 3.8) $R_n(t)$: sine response of harmonic n. $w_n(t)$: weight of sine response nth harmonic.
Characteristic Period Correlation (CPC): **$r(s,t,T^c(s))=c_s\ L\{ x_s(t)\ x_s(t+T^c(s)) \}$** Section 4.3 (see equation 4.9 for actual implementation).	$T^c(s)$: characteristic period for segment s c_s: normalizes CPC to energy of white noise Based on invariant structure of impulse response of TNC. Dominance if: $\frac{r(s, T^c(s), t)}{r(s, 0, t)} > 1 - C_C(s)$ $C_C(s) \ll 1$
Tuned Autocorrelation (TAC): **$r(s,t,T(t))=L\{ x_s(t)\ x_s(t+T_s(t)) \}$** Sections 2.4, 4.6 and 4.7	For given period contour $T(t)$ and segment group delay d_s: $T_s(t)=T(t+d_s)$. Application of periodicity as signal property: $r(s, t, T(t)) = \begin{cases} r(s, t, 0) & \text{Correlated } T(t) \\ \text{small} & \text{Uncorrelated } T(t) \end{cases}$
Running Autocorrelation along Ridges: **$r(s(t),t,t_2)=L\{ x_{s(t)}(t)\ x_{s(t)}(t+t_2) \}$** Sections 2.7 and 4.5	Ridges $s(t)$ arise through entrainment by periodic signal components. Strong ridges are dominated by a single signal component. Local approx. of interpeak distance gives accurate estimation of Local Instantaneous Frequency (LIF) development along ridge.

Table 7.1: Overview of representations and their properties. The notation has been generalized and differs slightly from the notation in previous chapters.

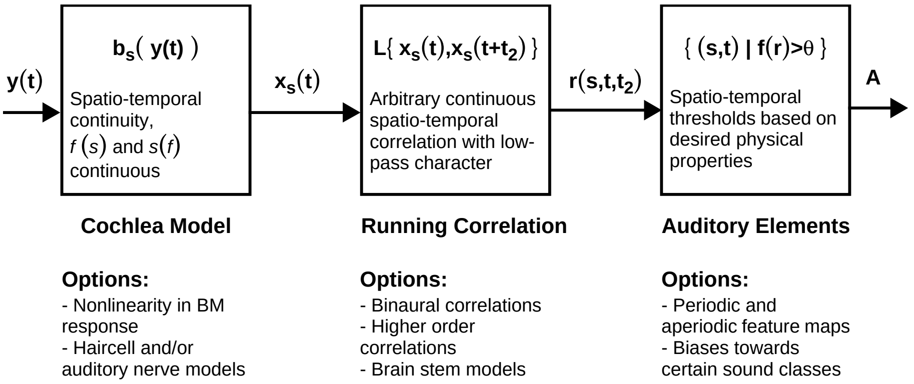
y(t)
b_s(y(t))
Spatio-temporal continuity, f (s) and s(f) continuous
x_s(t)
L{ x_s(t),x_s(t+t_2) }
Arbitrary continuous spatio-temporal correlation with low-pass character
r(s,t,t_2)
{ (s,t) | f(r)>θ }
Spatio-temporal thresholds based on desired physical properties
A
Cochlea Model
Running Correlation
Auditory Elements
Options:
- Nonlinearity in BM response
- Haircell and/or auditory nerve models
Options:
- Binaural correlations
- Higher order correlations
- Brain stem models
Options:
- Periodic and aperiodic feature maps
- Biases towards certain sound classes

Figure 7.2. System overview of *Continuity Preserving Signal Processing* (CPSP). An arbitrary input sound $y(t)$ is processed by a cochlea model $b_s(t)$. Running correlation operators $L\{\}$ compute special correlations $r(s,t,t_2)$. The final output of the system is set of auditory elements A that represent spatio-temporal points (s,t) with properties determined by a function $f(r)$ and a threshold θ. This scheme allows considerable freedom for optional features and alternative implementations.

frequency relation (such that $s(f)$ and $f(s)$ both exist) and the conservation of continuity through time and place. Optionally, the basilar membrane function can involve a *nonlinear compression* of the basilar membrane displacement as long as it conserves the temporal information related to entrainment and dominance. For example it is possible to apply physiologically plausible haircell and auditory nerve models (Siebert 1968, Duifhuis 1972, Meddis 1986). The implementation discussed in this work uses a linear BM model because linearity allows the analysis or reconstruction of cochleograms by superimposing suitably weighted harmonic templates (sections 3.3 and 3.4). The benefits of nonlinear models remain to be demonstrated, but nonlinear models might represent certain signal properties (typically forward masking related) better than linear models can (Duifhuis 1972, Tchorz 1999).

The running correlation operator $L\{\, x_s(t), x_s(t+t_2)\,\}$ must also preserve continuity through time and place. To be useful, L must respond slowly compared to $x(t)x(t+t_2)$ and consequently be slow compared to the characteristic period of the segment. In the current implementation, the lowpass filter is implemented as a leaky integration process with a fixed time-constant τ of 10 ms. In general the lowpass behavior of L may depend on the segment number s; for example to optimize temporal resolution. The use of τ=10 ms smoothens the correlations stemming from most segments sufficiently to warrant a sampling rate of the TNC (or derived representations) of 200 Hz.

Although the current implementation is limited to second order moments that stem from the same segments, this is not an essential restriction. In fact, $L\{\}$ may involve arbitrary correlations of the BM output $x_s(t)$. Since neurons can compute extremely complicated spatio-temporal correlations it is possible to model the running correlation in physiologically plausible (i.e., continuous time) neural networks. Inversely it might be possible to interpret the results of neurophysiological experiments in terms of CPSP. Binaural correlations are another obvious choice.

Different choices of the spatio-temporal correlations $r(s,t,t_2)$ represent different signal properties. Section 6.1 used the derived TNC subsets to form and identify *auditory elements*: areas of the place-time plane (s,t) that, most likely, represent the energy and frequency development of a *single* signal component. Because these regions are dominated by information of a single source, quasi-stationarity is justified along the ridges in these regions (theorem 1.21), which allows a correct estimation of the development of the instantaneous frequency. Sections 6.2 and 6.3 proved that these regions contained information conveying ridges (e.g., related to the temporal development of formants) that can be estimated in adverse signal-to-noise ratios. This fulfilled the tasks formulated in section 1.8.

Table 7.2 provides an overview of the criteria used to identify the auditory elements. The threshold values must be permissive: the strength of the approach lies in the combination of knowledge sources and not in attempts to optimize each of the knowledge sources separately. Auditory element estimation allows the reliable estimation of signal components at BM positions where the local SNR is 0 dB or better (section 3.6). A high local SNR allows a more reliable estimation of signal properties than a low local SNR. This links auditory element estimation to the experimental work of Fletcher, French Steinberg and Galt (French 1947) that showed that the local SNR is the main determinant of speech intelligibility (reviewed in Allen 1994). A bias towards speech is introduced by the choice of the minimal duration (e.g., 30 ms) of cochleogram contributions during mask forming.

The current system parameters makes the system especially sensitive to sound events that *change slowly enough to allow feature estimation in noise and fast enough to convey information*: this includes speech, animal vocalizations, music. etcetera. The next section proposes a method to use knowledge about the characteristic constraints of a target class (in this case speech) to build

Selection criteria	*Function*
Background model $A_B = \{(s, t) \| r(s, 0, t) > C_B(s, t)\}$	Focuses the search to points of the place-time plane (*s*,*t*) with (nonlinearly-scaled) energies above a background model C_B(*s*,*t*).
Driving energy $A_D = \left\{ (s, t) \mid \dfrac{\frac{\partial}{\partial t} r(s, 0, t)}{r(s, 0, t)} > C_D s, \tau \right\}$	The requirement of sufficient driving energy prevents the selection of (*s*,*t*) points that are not actively driven by a signal component at time *t*.
Local dominance $A_C = \left\{ (s, t) \mid \dfrac{r(s, T^c, t)}{r(s, T, t)} > 1 - C_C(s) \right\}$	Local dominance focuses the search to signal components that entrain a region of the BM actively. These regions contain, for periodic excitations, at least a single ridge.
Compliance to a period contour $A_T = \left\{ (s, t) \mid \dfrac{r(s, T(t), t)}{r(s, 0, t)} > C_P(s) \right\}$	Focuses the search to place-time plane (*s*,*t*) that show a periodicity conform an estimated period contour *T*(*t*). This period contour is a source property.
CPC close to energy $A_N = \left\{ (s, t) \mid \dfrac{r(s, T^c, t)}{r(s, 0, t)} < 1 + C_A(s) \right\}$	Ensures, in combination with A_C, that accepted values are close to local energy. This is a requirement for the selection of strong aperiodic signal components.
Periodic signal contributions $A_P = A_D \cap A_C \cap A_T \cap A_B$	Ensures sufficient driving energy, dominance, compliance to an estimated period contour and an energy exceeding the background model.
Aperiodic signal contributions $A_A = A_D \cap A_C \cap A_B \cap \neg A_P \cap A_N$	Ensures sufficient driving energy, dominance, an energy exceeding the background model, the absence of periodic contributions ($A \neg_P$) and CPC-values close to energy (A_N, see equation 6.11).
Mask forming $M = m_{L, H}(A)$	Reduces the accepted cochleogram area *A* to coherent regions of length *L*, with size *H* holes filled. Represents a bias towards speech-like sounds.
Speech like contributions $M_S = m_{L, H}(A_A \cap A_P)$	Selects coherent cochleogram areas. Assumes (implicitly) that speech is either periodic or aperiodic (which is false, e.g., /Z/ and /W/).

Table 7.2: Overview of the auditory element estimation criteria of section 6.1. The notation has been generalized to comply with table 7.1. Most of the acceptance criteria are related to physical properties. The bias towards speech signal is limited to the choice of *L* and *H* during mask forming.

recognition systems that function whenever the system is able to assign sufficient information to a single r epresentation (theorem 1.15).

7.2 Obtaining Acceptable Recognition Results

Chapter 1 formulated a number of demands for a speech recognition system that can deal with arbitrary sounds and that functions as often as (physically) possible in arbitrary environments. Such systems are called general recognition system (definition 1.4) which:

- must be based on the most robust features in the signal;
- must search through the set of possible recognition results and produce the best recognition result that matches a subset of the estimated features *sufficiently* (conclusion 1.12);
- must combine recognition and selection to avoid the signal-in-noise-paradox (definition 1.5).

The signal-in-noise-paradox is a direct and *inevitable* consequence of treating selection and recognition as separate processes, which is the case in traditional HMM-based speech recognition systems. These systems rely on the basic assumption that preprocessing can reduce the influence of background noise *before* the signal is recognized. Since this assumption is not universally valid, it limits the application scope of HMM-based recognition systems to special situations: notably situations without background noise or situations where the background is known or can be derived by rule. Applications for more general (and more useful) acoustic situations require another approach. This section proposes the outline of such an approach in a way consistent with the conclusions of chapter 1.

Reliable speech features

Because of the limits that the measurement process poses on the reliability of features, it is generally impossible to detect and/or to estimate the values of individual features completely or even sufficiently reliably (conclusion 1.16). This led to the conclusion that (speech) recognition systems ought to be based on *feature hypotheses* (conclusion 1.17). Generally, higher level hypotheses (which are based on multiple hypotheses) are more reliable than lower level hypotheses. Consequently feature hypotheses must be interpreted in context (conclusion 1.18). Consequently:

(7.1) It is not the detection and estimation of a feature that is of central importance, but its functional contribution to reach a meaningful (or at least an acceptable) recognition result (see conclusion 1.13).

It is to be expected that auditory elements that are able to dominate a sizeable region of the place-time plane will lead to reliable features, which in turn provide the most important contribution to a correct recognition result. Vowels represent signal components that are particularly well suited to dominate certain regions of the BM. In the low-frequency range individual harmonics at formant positions are strong narrow band signals that are quite able to dominate in the presence of more energetic broadband noise. The same is true for harmonic complexes at formant positions in the high-frequency range. In this case the increase in bandwidth is compensated by a decrease in frequency resolution of the basilar membrane. Furthermore the tuned autocorrelation groups the constituting harmonics, which results in an efficient reduction of the uncertainty associated with the interpretation of the signal.

The ability to dominate locally makes individual harmonics and harmonic complexes detectable, but to carry information entails that a signal must change. It is important to realize that speech signals represent little information and consequently a low rate of *relevant* change. Individual phonemes last typically more than 40-50 ms (with an average of 100 ms) and represent only a few bits. For example the classification of vowels requires a few bits to classify the formant pattern (most languages use less than 12 different vowels which suggests that 4 to 5 bits should suffice).

Miller (1955) and Nicely showed that individual consonants also represented about 5 bits. They derived a set of 5 features that allowed the classification of the 16 most common consonants in English. Miller's set consists of the binary features of *voicing*, *nasality*, *affrication* (the presence of aperiodic signal contributions) and *duration* and a triple valued *place of articulation feature*. Miller showed that confusions between consonants were highly predictable because they often resulted from an estimation error in a single feature value. Voicing and nasality were found to be very robust to estimation errors (probably for reasons as described in the previous paragraphs). Affrication, duration and the place feature are primarily important for unvoiced consonants and were much less robust. The threshold for the place feature lies about 18 dB above the threshold for voicing or nasality, which entails that the place of articulation of consonants is also difficult to estimate in positive SNR's.[1] Miller's conclusions are consistent with expectations based on the properties of CPSP.

With a phoneme rate of 10 Hz and an average of 5 bits per phoneme, the information rate of speech is in the order of 50 bits per second (Rabiner 1993). Consequently, speech decoding requires an equivalent of 50 binary decisions per second. To match the computational complexity of the recognition system with this information-rate it is necessary to base the decoding process on very coarse decisions in terms of time and phonetic feature values. Given the duration of individual phonemes an integration window of minimally 40 ms is to be expected. This interval is consistent with conclusions based on experimental evidence summarized in Greenberg (1996).

Different phonetic features are characterized by different signal properties (O' Shaughnessy 2000). During a *vowel or voiced consonant* a pitch contour can be estimated and TAC-selection of a large fraction of the energy is possible. Moreover, the CPC along the ridges of the phoneme can be greater than unity, the ridge's cross-section resembles a sine response (section 3.3) and the LIF-contours along the ridges show a smooth development. *Formants* correlate with the position of the most energetic signal components. Harmonic complexes at formant positions show an amplitude modulation with a rate equal to the pitch contour, which can be estimated as an amplitude modulation of the energy along the associated ridges. *Nasalized consonants* are voiced as well, but show weaker and broader resonances than vowels. Nasals have a first formant typically around 250 Hz, the second formant (near 1100 Hz) is very weak due to a spectral zero, while the third formant, near 2200Hz, is stronger than the second. The on- and offset of nasalization is often characterized by detectable transients due to the start or stop of the contribution of the nasal and oral cavities (Gold 2000).

Affrication is signified by low energy aperiodic contributions in the TNC (see section 4.2 and 4.3). The highest CPC-values are near unity and variable, and the main energy is usually in the high-frequency range of cochleogram. Affrication of voiced phonemes leads to a broadening (for voiced affrication as in /z/) or absence of sine-response like shapes. *Plosives* are characterized by a rapid build-up (see section 4.4). *Unvoiced plosives* lead to an aperiodic

[1] The place of articulation is the easiest cue to read from the lips of a speaker (Miller 1955).

TNC, while the TNC of *voiced plosives* shows transients that develop into the periodic form. Plosives are often characterized by vertical transients in the cochleogram or the CPC. The associated energy of individual plosives is often low compared to vowels and the main energy is often in the high frequency range. *Non-plosive consonants* show a development to a stable form without or with minimal on- and offset transients. *Place-features* depend strongly on phonetic context. *Back fricatives* have higher energy than *front fricatives*. Place cues reside often in the transition between the consonant and the adjacent phone (e.g., in the onset of second and third formant) and differ per context.

Since most features can be estimated in different ways, the best way to detect a certain feature will depend on the interaction between noise and target. In many noisy situations it may be impossible to detect certain phonetic features. The system ought to be able to deal with this uncertainty.

Syllables

All natural languages use syllables (here defined as a sonorant, i.e., periodic, nucleus with optional leading and trailing consonants) as basic building blocks for words. Apart from carrying linguistic information, the central vowel (or, more general, the most sonorant nucleus) forms a robust "anchor" for the less robust consonant features:

(7.2) Only (consonant) feature hypotheses in the few hundred milliseconds before and after the central vowel might be part of a speech signal.

Furthermore, the formant structure of coherent quasiperiodic signal contributions can be selected, even in unfavorable SNR's, with a very low probability that the estimated formant information stems from multiple uncorrelated sources. These are strong arguments to favor syllable wide decisions over more local decisions and to use the syllable as a *starting point for speech recognition*.

Syllables of languages like English and Dutch have a standard phonetic structure of which the realization as sound can be described as:

```
syllable = [*] [uC] [vC] <V> [vC] [uC] [*]
```
(7.4)

The square brackets [ ] denote zero or more repetitions, the angle braces < > denote one or more repetitions. The [*] denotes an optional silence, [uC] an optional unvoiced consonant, [vC] an optional voiced consonant and <V> one or more vowels. Since voiced parts can be estimated with the highest reliability and each syllable of normal speech contains a most sonorant centre, one might center a description of speech sounds on the sonorant nucleus of the syllables. This argument complements the discussion in Greenberg (1995) to explain importance of syllable-sized units in speech.

For example, words like *six*[2] /S IH K S/, *strength* /S T R EH NG TH/and the two syllable words *waiting* /W EY T IH NG/and *wailing* /W EY L IH NG/might be decomposed as:

```
               S     K  S
six     = [*]  -  IH -  -  [*]
               S  T  R      TH
strength= [*]  -  -  =  E NG -  [*]
               W     T
waiting = [*]  W EY  -  IH NG [*]
               W
wailing = [*]  W EY  L  IH NG [*]                      (7.5)
```

The baseline symbols represents the robust voiced nucleus of each syllable. The dash (-) denotes an unvoiced phoneme, the symbols above each dash give the phonemes identity and denote a combination of aperiodic features. The W above the baseline W in *waiting* reflects that, although a /W/ is voiced, it represents aperiodic energy as well. The equality symbol (=) in *strength* denotes that the /R/ might be either voiced or unvoiced. The continuous voicing of *wailing* forms very robust distinguishing evidence to separate *wailing* from *waiting*.

Recognition strategy

As noted before (e.g., conclusion 1.19 and the last paragraph of section 2.8) the optimal output of a feature estimation stage is a set of feature hypotheses. These hypotheses can activate interpretation hypotheses that might meet an *acceptability* criterion (definition 1.8). Acceptability entails that the characteristic requirements of a hypothesized class have to be satisfied, which demands that the feature hypothesis must be evaluated *in the context of the hypothesized class.*

The acceptable recognition result that accounts for the largest fraction of the input (in a way consistent which the state and tasks of the recognition system) is likely to be the correct result. One way to implement these requirements for speech is to combine three functional stages:

1. *A signal analysis stage* such as the auditory element technique summarized in table 7.2, which leads to a set of auditory elements of which each is likely to represent information of a single source;

[2] Phonetic transcriptions in TIMIT notation.

2. *A feature selection stage.* Auditory elements activate matching syllable models that begin to search for supporting evidence (typically formant development and evidence of less robust aperiodic features). Conflicting evidence deactivates syllables rapidly. Syllables that do not deactivate themselves are marked *acceptable* (as in definition 1.8). Acceptable syllables activate all possible successive syllables, which in turn check whether or not their vocalic nuclei (and other estimable phonetic features) comply to the input.[3]
3. *A global decision mechanism.* This stage determines which acceptable sequence of accepted syllables and which interpretation of the associated word sequences (given the system's state and context) explains the data best.

Figure 7.3 shows these functional stages and the representations that form their in- and output. Each stage has an associated temporal scope. The temporal scope of the signal processing stage depends typically on the time required to reach a stable local representation that allows an accurate feature estimation. This time scope is typically 5 to 20 ms for high-frequency segments and up to 50 ms for the low-frequency segments. This temporal scope matches the minimal duration of phones and the associated phonetic features. The feature selection stage must have a wider temporal scope to allow the evaluation of features in the context of a syllable. This scope is typically 100-300 ms. The decision stage must be able to evaluate multiple word hypotheses and consequently operate on a rate associated with the duration of words in context, the associated scope is 500 ms or more.

To allow an efficient and successful search, it is important to use the most reliable bottom-up evidence to activate a set of syllable hypotheses that includes the correct syllable. Favorable SNR's allow the estimation of more conflicting evidence and lead to a rapid reduction of the search space. Unfavorable conditions prevent the estimation of less robust conflicting features and lead to a larger search space. Consequently, the role of higher order regularities (e.g., syntax, semantics, discourse information) becomes more prominent for disambiguation in unfavorable SNR's.

[3] The activation of this type of models tells the system that the process they stand for is present in the input. These models have a function similar to the schema's as described in Bregman 1990. This function is similar to some of the ideas behind Adaptive Resonance Theory (ART) (Grossberg 1982, Carpenter 1987a, b), which in turn inspired psycholinguistic theories like TRACE (McLelland 1986), COHORT (Marslen-Wilson 1980), and Shortlist (Norris 1994).

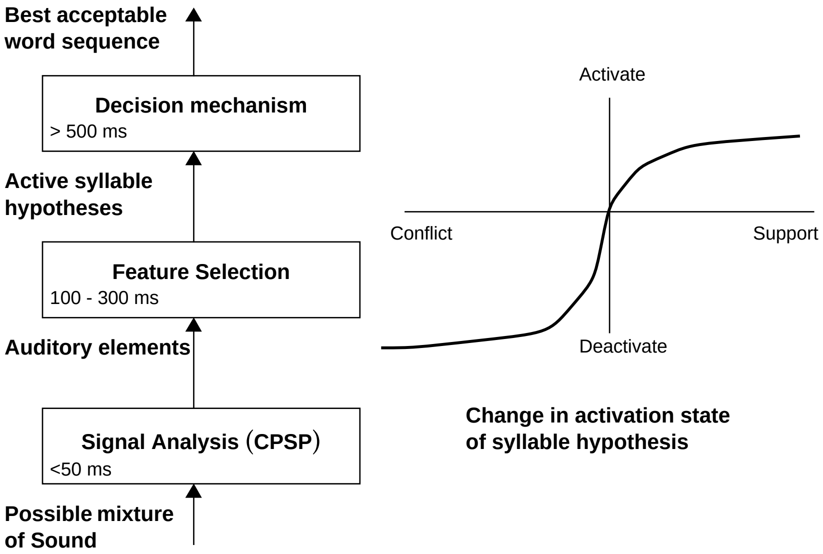

Figure 7.3. The proposed recognition system consist of three functional stages. The signal analysis stage (see figure 7.2) identifies auditory elements and analyses the associated energy and frequency content. The feature selection stage consists of syllable models that become activated by supporting bottom-up evidence and get quickly deactivated by conflicting bottom-up evidence (as depicted in the right-hand graph). The decision stage selects the optimal sequence of activated recognition hypotheses. The time indications provide an indication of the typical temporal scope associated with each functional stage.

Decoding

Although the task of a speech recognition system based on these three stages is different from standard HMM-based recognition systems, the decoding task is similar and can be applied with suitably adapted functionality. Figure 7.4 shows a part of a recognition network (adapted from Young 1996) that is used for *decoding* (i.e., estimating the best word sequence $\hat{w}$) in Large Vocabulary Recognition (LVR) systems. Because this network represents the set of all possible word sequences that the recognition system can deal with, it can be used to guide the search for acceptable recognition results.

The squares denote the words represented by the system (note that a silence is treated as a word). A sequence of activated words represents a recognition hypothesis. The other ovals are triphone models, e.g., B-EE+N denotes the model of an acoustic variant of the /EE/ that is preceded by a /B/ and followed by an /N/. The triphone models represent the acoustic constraints to

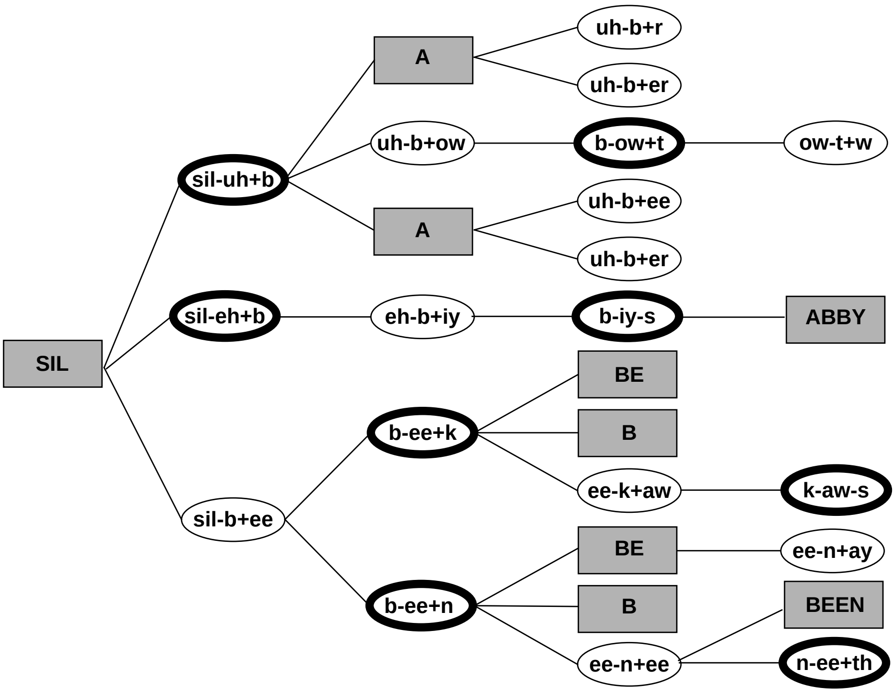

Figure 7.4. Part of a decoding network for large vocabulary speech recognition (based on Young 1996). The sequences of squares denote the word sequences that the network can recognize. The notation B-EE+N denotes a triphone model of an /EE/ in the context of a preceding /B/ and following /N/. The triphone models represent the acoustic constraints associated with the word sequences. The fat ovals denote the (robust) nuclei of syllables on which the decoding process is based. The other ovals correspond to consonants with less robust characterizing features that might not be estimable in certain noise conditions. The techniques in this work allow for a robust pattern matching where selection and recognition are integrated so that the signal-in-noise-paradox can be avoided.

be satisfied while a certain word sequence is pronounced. The fat ovals denote the most robust vocalic nuclei of syllables. These form the starting points of the feature selection process as described above. The triphone context provides additional, and more specific, expectations about the expected or required features. For example, the pronunciation of the central *t* in *Beat it* may change from an unvoiced /T/ to a voiced /D/ depending on whether or not the phrase is interrupted by a silence.

Speaker dependency

To separate concurrent speakers it is necessary to assign syllables, words and sentences to the correct speaker. This requires the use of speaker characteristics like spatial position (direction), average pitch, speaking rate, formant positions and dynamic features (such as timing) that are related to the articulatory dynamics of the speaker. Although each of these features can be ambiguous, combinations of these features are characteristic for individual speakers during a sentence. Speaker characteristics can be used to generate even more specific expectations that help to further r educe the search space.

When a new speaker starts, none of the speaker-specific expectations will fit, while more general, speaker independent expectations still match. In this case, a new speaker model must be derived from a comparison of the general model and the new data. While the interpretation of the sentence progresses, the new speaker model will become more specific and eventually evolve to a new specific speaker model.

Matching of expectations with model

Harmonics at formant positions form the most robust source of information in speech because they are the most energetic signal components. It is generally assumed that reliable formant information can be derived from the signal (e.g., O'Shaughnessy 2000, Young 1996). But a formant represents a *source* property *that can only be estimated from the signal when either the signal itself contains enough information, or when the signal can be interpreted in the context of a suitable expectation.* HMM-based systems that represent bottom-up estimations of the spectral envelope rely completely on the first approach. Especially for high-pitched speakers the shape of the first formant is often sampled by a single harmonic (for high-pitch children or soprano voices the first formant is often not sampled at all). This makes it very difficult to estimate the low-frequency region of the spectral envelope, and consequently the formant position, with any measure of certainty.

A more reliable way, consistent with the demand for *acceptability,* is to use the expected formant positions associated with a syllable model in combination with the estimated pitch to generate a detailed expectation of the cochleogram cross-section by the addition of suitably weighted sine responses as exemplified in figur e 3.8. An even more robust alternative is to focus on the ridges and to build ridge templates as in figur e 6.13. Section 6.3 showed that a ridge mask represents the most robust features and that a match, as defined in equation 6.15, between a suitably constructed ridge mask and the robust speech mask shows very little degradation for SNR’s above -5 dB (see figure 6.15).

When a suitable voiced phoneme model is constructed (e.g., as an average of Viterbi aligned examples of bottom-up estimated spectral envelopes), it is possible to search for evidence of strong harmonics at formant positions. If estimated, these serve as supporting evidence that enhances activation. Insufficient energy at the expected position or an unexpected temporal development indicates conflicting evidence, which deactivates the model. Finally, a situation with sufficient energy but an absence of estimable harmonics indicates that the expected evidence is likely to be masked.

This procedure is similar to the way HMMs try to reproduce the spectral envelope (see figure 1.3). In the proposed procedure however, the focus is not on the reproduction of a spectral envelope (which is likely to be corrupted due to unknown background noise), but on the explanation (i.e., prediction) of the development of those harmonics that represent the most important linguistic information and are least likely to be corrupted by noise.

Global optimization criteria

Any sequence of activated syllables and words represents an *acceptable recognition result*. The final recognition result is the *best acceptable recognition result*, i.e., a recognition result that satisfies *all* characteristic requirements at *all* levels of description. Generally, the best acceptable recognition result is the most meaningful interpretation of the data, but this is difficult to quantify. When *best* is defined as *most probable* it is possible to use the familiar *Bayes' decision rule framework* (Ney 1997) that dominates modern speech recognition approaches. This allows the application of equation 1.1:

$$\hat{w} = \underset{w}{\max \arg} \{P(y|w)P(w)\} \tag{7.6}$$

The acoustic models $P(y \mid w)$, that are usually implemented as HMMs, ought to be replaced by more versatile models (with the function as described above), but the language models $P(w)$ and the general decoding strategy and system integration can remain virtually identical.[4]

[4] A probabilistic framework is not the only possibility, other metrics are possible and may even be more suitable.

One of the weaknesses of HMM technology (Young 1997) is the statistical independence assumption that is used to justify the multiplication of probabilities during the recognition process. HMM system often use a 40-dimensional input vector.[5] When each of these values contributes with probability 0.5, the combined probability is $0.5^{40} \approx 10^{-12}$ per 10 ms frame. For a 1 second word the associated probability is $1/10^{1200}$, a value that is unrealistically low for a model that might correspond to the correct word in a ten digit task. This low probability is the result of the unjustified application of the independence assumption. Yet, as long as it can be guaranteed that the correct word sequence is still the most probable sequence, these low probability score do not necessarily lead to problems.

However, an insightful experiment by McAllaster (1999) shows that problems do arise even in situations without background noise. McAllaster compared the performance of an ASR system trained on spontaneous speech (from the Switchboard corpus) when presented with either real or fabricated speech data. The fabricated data was designed to comply perfectly to the demands of the correct model sequence, i.e., the fabricated data represented a *statistical independent* version of the real input data. The word error rate on the real data was 48.2% but on the fabricated data the error rate dropped to only 4.3%. This led McAllaster to conclude that *the failure to model* (*the varied pronunciations of spontaneous*) *speech properly is to blame for most of the errors of the ASR system*.

For improved speech models *M*, one might require that the acoustic model $M(y|w)$ ought to produce a number, termed *quality*, reflecting the "probability" that a given syllable or word (and not the individual elements of the input vectors) matches the available acoustic information. For example, whenever a dialogue system expects the word *yes* and the recognition system determines that the characteristic requirements of the word *yes* have been satisfied, the produced quality should be close to 1. This couples the *quality of model M* to a measure of acceptability. Acceptable recognition results receive a quality close to 1. If the actual input was *yep* some of the less important constraints are violated and the model for *yes* must receive a low quality. Unacceptable recognition results ought to have a very low quality and are discarded from the search space. Note that when the actual input was *no*, the model for *yes* should not have been activated in the first place.

[5] 12 MFCC coefficients plus log energy and their first and second order temporal derivatives lead to 39 values to describe a single 10 ms input frame.

Recalling that each phoneme represents only a few (e.g., 5) bits, one can assign the quality 1 to any feature that has been estimated within an acceptable range. Features that are not within the expected range ought to receive a quality reflecting the deviation from the expectation, the feature's importance in reaching correct results and a measure of the probability that an estimation or speech production error occurred. Important features that are based on reliable information, but are well out of range are considered as conflicting information that deactivates the word and eventually removes the hypothesis from the search space. In noisy situations, some of the less robust features may not be estimable because the expected energies are below the background level. In these undecidable situations a (*"don't care"*) quality of 1 ought to be assigned to the feature.

This scheme can only produce relatively high quality recognition results because low quality results are eliminated from the search space and the remaining hypotheses are all *acceptable* given the estimable information. The statistical independence assumption is neither used nor necessary because each feature is evaluated in the context of a syllable hypothesis that takes care of the necessary dependencies.

It is possible that multiple acceptable recognition results are produced with quality 1. Take for example the word *stop* produced in silence. A recognition system as proposed in this section ought to produce the correct result *stop*, but might also produce the word *top* with quality 1 because all characteristic requirements are satisfied. The language model can help to make the correct choice, but, alternatively, the *activation state* of the two words can be taken into account. The word *stop* expected and found evidence for an /S/. This supporting evidence increased its activation above the activation of *top* that did neither expect nor require additional information. A suitable combination of recognition quality and the associated activation state of *all* words can be normalized to add-up to unity and to yield a probability $P(y|w)$ that an acceptable word sequence w is the true word sequence given the acoustic information y. Together with a standard language model P(w), the familiar Bayes' decision rule framework of equation 7.6 can be applied as usual.

Decoding efficiency

The decoding strategy can be implemented efficiently. Compared to the traditional recognition approach, more emphasis is placed on the preprocessing stage, which produces auditory elements that may, or may not activate models of syllable nuclei. In situations without speech-like signal components, no further computing cycles are wasted on attempts to match the signal with word models. Activated syllable nuclei search within a few hundred millisecond wide scope of auditory elements for fairly coarse grained evidence (Greenberg 1996), which can be performed with a low computational load, spatio-temporal points without local energy or points where the local energy exceeds the energy of the expected features can be ignored.

Activated words with a dense neighborhood (i.e., a relatively large number of similar sounding alternatives) may require a more careful analysis to determine the most likely choice. For example the sentences *To recognize speech* and *To wreck a nice beach* are very similar acoustically and might become active concurrently with very similar qualities. Yet, when the system knows that *speech* and *beach* are easily confused, it is possible to analyze the phonetic features of the complete sentence hypothesis better. In this case, intonation pattern differences and a minimal affrication during the /W/ and an early voice onset time during the /B/ might favor the last sentence over the first.

To reduce computational load of the recognition system even further, it is important to organize the search efficiently. This requires that the characteristic constraints of the expected and/or high-frequency words (or syllables) are checked before unexpected and/or low-frequency words. Furthermore it is important to make an inventory of confusable syllables and words. The first (expected) acceptable recognition result that is unlikely to be confused with other word sequences signals a correct result and the end of the search.

Out of vocabulary words can be detected in sentences where at some intervals syllable-level constraints are satisfied while a valid wor d cannot be formed.

Psycholinguistic relevance

This section proposed a recognition strategy based on the recognition framework of chapter 1 and the techniques developed in later chapters. It is argued that more suitable signal processing enables a very efficient decoding scheme that can be used to recognize sentences of individual speakers in arbitrary environments (the cocktail party effect, Cherry 1953). This strategy is consistent with the solution (actually avoidance) of the *signal-in-noise-paradox* that was proposed in conclusion 1.12. The strategy is also consistent with some modern insights of speech processing (Greenberg 1995).

A well know effect is *phonemic restoration* (Warren 1970, Samuel 1996): this involves the filling-in of one or more phonemes (even without any acoustic evidence) whenever sufficient masking energy exists at positions of the time-frequency plane where phonetics evidence is required to reach a certain recognition result. The actual phoneme that is filled-in may depend on the linguistic context and may change when more linguistic information is provided (Samuel 1996). This is consistent with the recognition strategy proposed here. Other experiments have shown that the reaction times associated with a detection of the leading consonants are longer than the reaction time associated with the detection of the syllables (Connine 1996). This is consistent with a (human) decoding strategy based on syllables.

7.3 Conclusions

This work is based on two conjectures. The first conjecture entailed that to be optimally useful, the human auditory system needs to function as often as possible in variable and unknown acoustic environments. The second conjecture stated that the most informative linguistic features were also the most robust features. This allowed a focus on the identification and estimation of (robust) features that can be used for recognition, resynthesis and further analysis. Care has been taken to base design decisions on the functional requirements of recognition systems that function in as many acoustic environments as possible.

Since it is not certain that the conjectures hold for the human auditory system, explicit design decisions based on psychophysical or psycholinguistic evidence have been avoided. Yet the fact that some important psychophysical results can be explained in terms of CPSP suggests that the conjectures lead to results that are consistent with some important properties of the human auditory system.

- *Focus on local SNR.* Section 3.6 concluded that, given a correct period contour, the local SNR is the main determinant of the quality of the tuned autocorrelation. Section 6.1 generalized this by identifying and selecting auditory elements as time-place regions that are likely to be dominated by a single source, consequently these regions will show a positive SNR for that source. This linked the developed techniques to the experimental and theoretical work of Fletcher, French, Steinberg and Galt (French 1947, reviewed in Allen 1993) that showed that *the local SNR and not the spectrum is the main determinant of the intelligibility of* (*nonsense*) *words.*

- *Accuracy of frequency estimation.* Section 2.7 and section 4.5 demonstrated that an accurate local frequency estimation is possible. The accuracy is in the same range as human performance (less than 1% at 1000 Hz). *This is only possible by avoiding the time-frequency trade-off of frame-based approaches and forms important supporting evidence for the importance of the conservation of continuity.*
- *Breakdown of performance below 0 dB.* Section 7.3 argued that as long as a minimal set of informative auditory elements exist that activate bottom-up expectations and deactivate the incorrect expectations, it is possible to recognize speech with a low probability of error. The robust speech masks of section 6.3 represent only a small fraction of the total area of the time-place plane, yet they represent almost all linguistic evidence. Given the small area of the robust speech mask, its main features can only be masked by uncorrelated noise in SNR's that are well under 0 dB. This is consistent with *Speech Reception Threshold* (SRT) experiments (Plomp 1979, Alefs 1999) that show that, in common broadband noises, at around -6 dB 50% of meaningful sentences is not correctly recognized. Given a sentence length of 10 words this corresponds to a recognition error probability of 5%. An implementation of the proposed recognition strategy that leads to a similar error-rate proves the consistency of the basic conjectures with human performance.

These phenomena constitute important evidence that the application of conjectures 1.1 and 1.2 leads to results consistent with the human auditory system. Consequently CPSP may lead to ASR systems with *human quality performance*.

This work introduces a novel framework for the analysis of time-varying signals with certain properties:

- The approach is based on very weak assumptions about the signals to be preprocessed: consequently it is able to deal with varying, uncontrollable, unknown and arbitrary conditions.
- If, as is claimed, the basic assumptions are indeed maximally weak, than this form of sound processing is optimal for the analysis of unknown arbitrary signals.
- In its current form, the techniques are especially suitable for the analysis of (natural) information carrying signals where the information rate is maximal given the requirement that the features that carry the information must be estimable in a maximally wide range of acoustic situations. Put differently: *these techniques are suitable for the study of natural signals that are*

optimized for maximal communicative success in variable environments. Typical examples of this class of signals include speech, music and animal vocalizations.

- CPSP can track the development of individual physical processes that produce complex signals in a varying acoustic background. Apart from speech processing this suggests potential applications as diverse as EEG- or ECG-analysis, process control by sound and auditory scene analysis.
- The framework avoids an a priori decision about the time-frequency trade-off that is characteristic of frame-based methods. Consequently it allows a very accurate estimation of frequency content as well as the estimation of temporal detail.
- The formalism is not based on mathematical convenience, but inspired by physics, physical measurement theory and very general optimization criteria, this may be an intuitive approach for some users of CPSP.
- CPSP allows the separation of aperiodic and periodic components of sounds. This property might lead to interesting new analysis tools for the analysis of physical sound sources.
- The mask forming technique of CPSP (section 6.1) can lead to an improvement in the SNR of up to 20 dB (section 6.3).
- The place-frequency relation can be warped to suit the demands of the task. As long as the place-frequency relation is invertible, it is possible to study different frequency regions in varying detail (e.g., to create an auditory fovea as bats have for active sonar). Note that an increase in spectral resolution is balanced by a matching decrease in temporal resolution due to an increase in group delay (equation 4.13).
- At this moment it is possible to compute the complete analysis from the BM model up to auditory element estimation in real-time for 8 kHz signals on a 1 GHz computer. The BM model requires an important fraction of the computational load, but considerable reductions of computational load are to be expected. This will enable the application of CPSP in lightweight, low-power applications like mobile phones.