CHAPTER 1

Recognizing
Arbitrary Sounds

This chapter provides a framework for speech signal processing that is intended to function in arbitrary, varying, uncontrollable and unknown acoustic environments. The framework is based on insights from both physics and cognitive science, which is in contrast with common speech signal processing techniques that rely on mathematics and typically use standard electrical engineering techniques based on digital signal processing and applied statistics. The new approach is based upon a number of straightforward basic design considerations that are presented in this chapter. Subsequent chapters explore the consequences of these design considerations.

1.1 Speech and Speech Recognition

What is speech, and what distinguishes speech sounds from other sounds? These are hard questions because speech sounds are meaningful sounds. One might consider speech as an intermediate product of a very complex system that can communicate conscious thoughts and ideas of a speaker to a listener. In the case of speech the intermediate product is meaningful sound, in the similar case of sign language the physical manifestation is a set of meaningful gestures. In both cases, it is impossible to determine the meaningfulness of the
physical manifestation by studying it isolated from the rest of the system. Just like most people are unable to determine whether or not a combination of gestures is part of sign language, so is even the most competent linguist unable to determine whether or not a sound is speech when he or she does not know the language.

Like human listeners, an automatic speech recognition (ASR) system cannot determine that a signal is speech, unless the system knows about the regularities of speech sounds, the regularities of the language and the structures of the world that are associated with meaning. When the system does not know enough of the structure of speech sounds, it will treat nonspeech sounds as speech or, inversely, it will ignore valid speech. When the system does not know how speech sounds ought to be combined to form valid sounds of e.g., English or Dutch, it will treat foreign or nonsense words as meaningful words of the target language. Vice versa it might ignore or misinterpret valid words. And when the system does not know enough about syntax and semantics it is likely to choose a nonsensical combination of words as recognition result.

Building a speech recognition system is therefore an extremely challenging task: it requires the integration of signal processing, linguistics and knowledge of the world. It requires in fact much more knowledge of these fields than is currently available. Designing a speech recognition system with the competence of a normal human listener is therefore beyond our current capabilities. At this moment ASR systems cannot reliably recognize spontaneous speech or speech with a moderate amount of background noise that does not impair human performance. Nevertheless, technology allows us to build systems that function reasonably well in one important special case: namely when the user ensures that the system’s input is limited to sounds the system can process correctly. When the user cannot or does not satisfy this condition the system will fail and/or produce a nonsensical output.

An important example of a useful speech recognition application is a dictation system. Such a system requires the user to train it:¹ for optimal results hours or even days of training are required. It is speaker-dependent in the sense that it functions well with the user(s) that trained the system, but functions suboptimally (or fails) with all other users. Yet, these systems, when functioning, allow the users to interact with their computer in a very natural

¹ At the same time the user is trained to produce speech in a manner that the ASR system can recognize.
way. Unfortunately not many users are willing to spend the necessary hours on training. Hitherto these systems are more sold than used.

Another useful application area of speech recognition technology is that of automatic telephone information services. These information systems function best when the user is cooperative and experienced and when the dialogue is very structured. A structured dialogue reduces the number of possible words, which facilitates recognition enormously. Because the dialogue is very structured and the system usually does not need to recognize each word to estimate the next course of action, these systems can function satisfactorily for a large number of cooperative users.

Before addressing the basics of modern speech recognition systems, it is useful to describe some aspects of the speech signal (O'Shaughnessy 2000, Gold 2000). Speech is produced in the vocal tract. Its energy stems either from the flow of air around constrictions in the vocal tract, which leads to a signal in which all frequencies contribute (a continuous frequency distribution). Or it stems from the periodic opening and closing of the vocal folds in the glottis; this restricts the signal to a discrete set of frequencies that are periodic with the main period of the vocal fold movement. The resulting sound is changed (filtered) by the resonant properties of the vocal tract. The excitation of the vocal folds determines the pitch and the loudness, which carry intonation and information about speaker identity. Pitch and loudness are relatively unimportant for meaning in most European languages. The filtering effect of the vocal tract, on the other hand, is the main source of information for word-identity estimation.

The speech production process is often modeled by a source-filter model as depicted in figure 1.1. This figure provides two equivalent versions of this process. The upper panel shows a time-domain representation in which the driving force is either a periodic pulse produced by the vocal folds, or an aperiodic contribution from turbulent flow. This excitation is modulated by the resonance effect of the oral, nasal, and pharyngeal cavity. The resulting speech is radiated through lips and nose.

The lower panel represents the same process in terms of the frequency domain. Now the vocal fold excitation is a set of harmonics: frequency components at integer multiples of the fundamental frequency ($f_0$) of the vocal fold movement. The first harmonic, $h_1$, represents the lowest frequency of the vocal fold movement and equals $f_0$, the nth-harmonic, $h_n$, represents $n$ times the fundamental frequency. This discrete set of frequency components is
Recognizing Arbitrarily Sounds

Continuity Preserving Signal Processing

Filter ed by the vocal tract. This leads to a relative enhancement of some of the harmonics, and to a relative attenuation of others. Frequency regions where harmonics are enhanced are termed formants. The first two formants, \( F_1 \) and \( F_2 \), are most important for word-identity estimation. \( F_3 \) is somewhat less important. Higher formants are of no special importance for word-identity estimation, but do contribute to speaker identification.

The upper panel of figure 1.2 depicts the waveform of the Dutch utterance \(/\text{NUU EEN TWEE DRIE}/\) (English translation: \(/\text{ZERO ONE TWO THREE}/\)). Because the temporal resolution of the picture is low, it is impossible to identify individual oscillations. But even with a good temporal resolution the
waveform is hard to analyze. By performing a so-called Short Term Fourier Transformation (STFT), the signal can be transformed to a frequency domain representation that is easier to analyze. The STFT divides the signal into a number of frequency contributions, or spectrum, that describe the signal during a certain time interval. These intervals are called frames or windows. A succession of spectra, describing the temporal development of a signal, is called a spectrogram or time-varying spectrum (Papoulis 1984). An example is depicted in the lower panel of figure 1.2. The horizontal axis represents time, the vertical axis frequency and the color coding reflects the relative contribution of a certain frequency within a certain time window: blue for low values, yellow for intermediate values and (dark) red for the highest values. Spectrograms are generally believed to provide sufficient and suitably organized information for speech recognition purposes (Furui 1989, O’Shaughnessy 2000, Young 1997).

The individual harmonics show up as slowly varying (horizontal) structures. Natural sources and in particular speech are rarely perfectly periodic: usually
Recognizing Arbitrary Sounds

### Template database

\[
\{ \{'1', '2', '3', '4', '5', '6', '7', '8', '9', '0'\} \}
\]

**Combinations** | **Probabilities**
--- | ---
'1' | P('1')
'11' | P('11')
'111' | P('111')
... | ...
'12' | P('12')
'121' | P('121')
'1211' | P('1211')
... | ...
'0008' | P('0008')
'0009' | P('0009')
'0000' | P('0000')

\[
w = \max_w \frac{P(w|y)}{P(w)} = \max_w \frac{P(y|w)P(w)}{P(y)}
\]

![Diagram](image)

Figure 1.3. The standard speech recognition paradigm is based on Bayes’ decision rule which is depicted in the upper right corner. The input is assumed to consist of words for which probabilistic templates (Hidden Markov Models, HMMs) exist. The combination of templates with the highest probability to reproduce the input signal is the recognition result.

the fundamental frequency, or its perceptive analogue *pitch*, changes as a function of time. This is reflected in the frequencies of individual harmonics that must remain integer multiples of the fundamental frequency. Signals with changing pitch are called *quasiperiodic* and form the main study object of this work. Formants correlate to areas in the time-frequency plane where harmonics are particularly prominent. Between \( t=0.2 \) and \( t=0.3 \) s a formant glide can be observed. During this period the second formant drops from 2000 Hz to 700 Hz, which is caused by a change of the vocal tract during the /L/ of /NUL/ that results in a matching change in the relative contribution of individual harmonics.\(^2\) Most speech recognition systems use the Short Term Fourier Transformation as the basis for further analysis.

Modern automatic speech recognition (ASR) systems are based on *Bayes’ decision rule* as depicted in figure 1.3 to determine which word, or a string of words corresponds to a given sound. Bayes’ decision rule leads to a probabilistic framework where each sound is modeled as a so-called Hidden Markov Model (HMM). Each HMM is specialized to produce a word or

\(^2\) Note that the harmonics go up, while the formant goes down: changes in formants and changes in pitch are uncorrelated.
phoneme in the sense that it has a relatively high probability of producing the corresponding word or phoneme and a much lower probability of producing other words or phonemes. The probability that a string of word templates \( w \) (or other unit of speech) produces the observation sequence \( y \) is denoted as \( P(y \mid w) \). Some word strings are more probable than others. This is denoted by the stochastic language model\(^3\) that is represented by a set of values for \( P(w) \).

The word string \( w \) with the highest probability of producing \( y \), forms the output of the recognition system. This process finds the best word sequence \( \hat{w} \) given input \( y \), i.e., it maximizes \( P(w \mid y) \), according to Bayes’ decision rule:

\[
\hat{w} = \max_w \arg \frac{P(y \mid w)P(w)}{P(y)} = \max_w \arg \frac{P(y \mid w)}{P(y)}
\]

The second step on the right side results from the application of Bayes’ rule. \( P(y) \) is independent on the word sequence \( w \). Automatic speech recognition systems require a suitable form of the input\(^4\) \( y \). This input is usually based on a spectrogram-like representation that is tailored to estimate the spectral envelope as efficiently as possible while minimizing pitch information in the parameterization. Typical examples are Mel-Frequency scaled Cepstral Coefficients (Young 1997, Gold 2000) and Perceptual Linear Prediction (Hermanski 1990). Both represent approximations of aspects of human perception, but do not make any difference between the type of input: consequently a speech signal is processed in the same way as a nonspeech signal or a mixture of speech and noise. This limits the application of speech-specific knowledge.

\(^3\) A stochastic language model of order N represents the probability of a word given its N-1 predecessors. Syntax, semantics, and other linguistic knowledge is implicitly coded in these N-grams.

\(^4\) To simplify implementation and to speed-up processing, most ASR systems assume statistical independence of consecutive (overlapping) input vectors and statistical independence of the parameter values that constitute each input vector. This assumption is equivalent to assuming that speech is produced by a random process where each parameter of the input is uncorrelated with all other parameter values. Given the speech production overview in figure 1.1 and the discussion in section 1.7, this assumption cannot be justified.
The knowledge about the development of speech sounds is stored in the acoustic models \( P(y|\omega) \). These models are based on as much speech data as possible. After “training”, each model represents the corpus mean (and variance) of the sounds it represents. Consequently, HMM-based recognition systems represent a form of pattern matching where the input is compared with the average development of the training corpus. Often separate models are trained for male and female voices. Although this recognition approach was not originally intended for noisy speech, it is used as a natural starting point for the development of truly noise robust applications.

1.2 Research Approach

This chapter started with the observation that speech is meaningful sound and that its meaningfulness cannot be studied in isolation from the rest of the (communication) system. Since the rest of the communication system involves large parts of our brain and the articulatory and auditory organs, the speech signal itself might only be the top of the iceberg in the sense that most of the regularities in the signal are due to processes and limitations of the rest of the system. When this is the case, the speech signal cannot be understood adequately without some understanding of the basic processes that produce and process it. It is even possible that the optimal (and natural) representation of speech depends on the possibilities and limitations of the whole cognitive system. A suboptimal choice of the basic representation of speech and the required functionality to process speech might be part of the reason why it is still impossible to develop reliable ASR systems for spontaneous and/or noisy speech.

One might therefore expect that ASR research necessarily involves an integration of (cognitive) science and technology. It does not. The development of ASR techniques in the last decades was almost exclusively technology driven and not science driven. On the other hand, research in the diverse fields of cognitive science (the sciences that study the algorithms of the brain) has been extremely fragmented and often aimed at explaining experimental data obtained by studying idealized stimuli in controlled environments.

The history of ASR technology (Gold 2000) shows that ASR technology has converged to the widely accepted HMM approach. There are some alternative schemes, usually involving artificial neural networks, which may lead to systems that function just as well. But these systems suffer from the same basic
limitations as HMM-based systems do: they function well within a certain range of operating conditions and fail when these conditions are not met.

Engineers are educated to select, use, and adapt a subset of the large set of well known and extremely powerful engineering tools. When years of research fail to produce systems with the desired functionality one might consider the possibility that the current set of engineering tools is not suitable to deal with the problem at hand. In fact, it suggests a problem with the basic assumptions on which modern ASR systems are based. The set of basic assumptions comprises all that the system takes for infallible: consequently when these assumptions are not justified, the systems will produce incorrect (and even unpredictable) results.

This behavior is in stark contrast to the reliability of human speech recognition performance. This leads to the question: which set of basic assumptions allows such a reliable human speech recognition performance? One might expect this to be a core question of the cognitive sciences that study the speech process. It is not. Cognitive science is still very fragmented and in search of a common general framework. Consequently it has great difficulty in answering questions about the performance of the whole (speech) process/system. Most cognitive scientists that study speech, study abstractions of the speech process that are still far from conversational speech in a normal acoustic environment. This results in scientific modelling that is either too abstract to implement (and therefore very difficult to falsify on real-life situations), or too detailed and complex to be of instant use in any feasible engineering system.

The approach of an engineer is to build a system that works with a subset of standard engineering techniques. The approach of a cognitive scientist, on the other hand, is to understand aspects of cognition with the learned conceptual background and the experimental techniques. The focus of an engineer is aimed at building a working and integrated system, while the focus of the cognitive scientist is (often, but not necessarily) aimed at the understanding of the processes that make up the system. This suggests that the approach of the engineer and the cognitive scientist might complement each other. This work tries to find a hybrid approach by identifying and applying aspects of cognition that allow the implementation of more reliable speech signal analysis systems.
1.3 Defining a Speech Recognition System

As a first step towards the formulation of a set of reliable basic assumptions it is useful to formulate the criterion of a speech recognition system. This is what the developer tries to achieve and consequently what is optimized in the system. In ASR-research the criterion is traditionally, the reduction of word-error rate (typically by improving the acoustic model \( P(w|y) \)) or the increase in the fraction of successful dialogues (typically by improving the language model \( P(w) \) and the system integration). These measures are relevant for human speech recognition as well. But human perception, being the product of evolution, optimizes behavioral benefit in combination with the requirements of the environment. One can therefore postulate that the human auditory system balances requirements like:

- quality of recognition,
- speed of recognition,
- the ability to rapidly and correctly process unexpected and potentially behaviorally significant sounds,
- the number of (acoustic) environments in which the system works adequately,
- source characterization (e.g., speaker identification)
- the integration of perception with behavior (the effect of the rest of cognition).

Human behavior shows that words in sentences are recognized almost flawlessly within 1 second (typically 600 to 800 ms), in a wide range of signal-to-noise ratios and in a wide range of qualitatively different acoustic environments (Alefs 1999). More difficult acoustic environments require longer processing (up to a few seconds) and more attention. In most acoustic environments the transition, in terms of the signal-to-noise ratio, from unimpaired speech perception to the complete inability to recognize speech is usually sharp. Around the speech reception threshold (Plomp 1979) the fraction of sentences of which all words are recognized correctly, decreases typically with 15% to 20% per dB signal-to-noise ratio. For a noise condition with a speech reception threshold of -6 dB, all sentences are correctly recognized above -2 dB, and none of the sentences is recognized correctly below -10 dB.

The wide range of unimpaired recognition performance in combination with the sharpness of the transition between unimpaired performance and the complete inability to recognize speech suggests that: human speech perception is
Defining a Speech Recognition System

Continuity Preserving Signal Processing

optimized to function, both reliably and as fast as possible over a range of signal-to-noise ratios that is as wide as possible and in a number of realistic acoustic environments that is as large as possible. Two tentative conclusions can be drawn from this observation.

The first conclusion is that the human auditory system optimizes the fraction of time (assuming real-life conditions) that it functions adequately. This forms the first basic conjecture of this thesis:

(1.1) The human auditory system is optimized to function as often as possible in varying and uncontrollable acoustic environments.

In ASR research the term robustness is used to denote a similar property: a method or algorithm is robust if it can deal with a broad range of applications and adapt to unknown conditions (Junqua 1996).

The second tentative conclusion is that the speech communication process is optimized to function as often as possible in varying and uncontrollable acoustic environments. This leads to the second basic conjecture of this work:

(1.2) The most important (linguistic) information in a speech signal is represented by those signal properties that can be estimated by the human auditory system in the widest possible range of real-life conditions. These signal properties are, by definition, the most robust properties.

Although this conjecture may seem obvious to some, it is actually difficult to prove because the information in a speech signal cannot be quantified since a quantitative theory of linguistic meaning does not exist. Yet speech communication, by definition, requires the transmission of information. This second conjecture states that the signal contributions that can be estimated in the widest range of (real-life) environments are also the most informative. This allows a change of the search effort from optimally informative linguistic features (which is difficult to quantify) to a search for features that can be estimated in a maximally wide range of acoustic environments (which is easier to quantify).

Although conjecture 1.2 is difficult to prove, there is ample supporting evidence. As discussed in the context of the figure 1.1, the peaks in the energy spectrum, which are likely to exert the largest influence in a measurement system, represent the formant information that is associated with word identity. ASR systems use formant information by modelling the spectral envelope. Speech synthesis systems (Keller 1994) use the formant development as the main determinant of intelligibility, and speech-like signals that consist of sinusoids at each formant position are intelligible for trained listeners (Remez 1981). Furthermore, work in the period between 1921 and
1951 by Fletcher, French, Steinberg, and Galt (French 1947, reviewed in Allen 1993) showed that intelligibility scores can be predicted by the articulation index: a measure related to the signal-to-noise ratio (SNR) per frequency band. This suggests again an important role for spectral peaks, because they correspond to positions in the time frequency plane where the signal-to-noise ratio is maximal. Consequently, spectral peaks are likely to carry linguistic information.

Note that conjecture 1.1 can be rephrased by stating that the human auditory system is able to correctly process arbitrary sounds as often as possible. The term arbitrary sound now has a special meaning:

(1.3) An arbitrary sound is a sound of which no a priori knowledge is available: it can be any sound out of the set of all possible sound combinations.

This means that of an arbitrary sound nothing is known, except for the most general properties that are valid for all sounds. An arbitrary signal might, or might not be recognizable by a recognition system. Modern ASR system cannot recognize arbitrary sounds, and not even arbitrary speech sounds. The correct functioning of an ASR system is usually restricted to a very specific set of words, speakers, types of background noises, transmission and coding properties etc. This entails that, in the context of ASR, an unknown sound has an extremely restricted meaning; virtually everything of the signal has to be known and conform to the system’s expectations; the only aspect that is left to be estimated is the word order. When the signal does not comply to the expectations of the system, it fails. Such systems are speech recognition systems in the sense that they recognize any input as being speech.

A much more useful system is able to deal with arbitrary input. This can be termed a general recognition system:

(1.4) A general recognition system is a system that can recognize a signal that belongs to a certain target class, if and only if it is actually present in the input.

Note that the term general does not refer to the recognition of all types of input, but is aimed at recognizing all signals of a certain class embedded in an arbitrary background. The rest of this chapter presents a framework for the development of general (speech) recognition systems. Chapter 7 returns to this framework and reflects on the way the techniques developed in this work help to design general recognition systems.

---

5. This led Allen (Allen 2000) to conclude that the role of the formant is to improve the local SNR.
The Signal-in-Noise-Paradox

1.4 The Signal-in-Noise-Paradox

This section is devoted to a more careful presentation of one of the main problem one encounters while recognizing arbitrary sounds: how to decide which part of the sound is the target or otherwise meaningful without having to recognize the signal. This problem becomes apparent while studying a single example of a signal that might be recorded at a lively cocktail party. A spectrogram of such a signal\(^6\) is depicted in figure 1.4. Suppose we aim to recognize speech: that means that we are not interested in the unintelligible background babble or in nonspeech contributions like music. We just want to recognize any intelligible speech that is part of the signal. But since we have not recognized the signal, we do not know where that speech is, what is said

\(^6\) Available at http://www.bcn.rug.nl/andringa/thesis

Figure 1.4. A Fourier-based spectrogram of a recognizable signal in cocktail party noise. It is difficult to find all signal components that belong to a single source. This figure was derived with a standard FFT-spectrum based on 25 ms Hamming windows and a time step between consecutive frames of 10 ms. A logarithmic energy compressing was applied. The signal-to-noise ratio is 0 dB.
Recognizing Arbitrary Sounds

or who was speaking. In fact, usually we even cannot be sure that there is any speech at all. Since we do not know what to expect, we do not know what to look for. In this example the situation is even more difficult because most of the signal consists of speech sounds.

Figure 1.4 shows a lot of structure: part of the visible structures reflects the target speech, part the background and part a mixture of both. If we could select the target speech, we could discard the background and present the selection to an automatic speech recognition system. If we made the correct selection, the probability that the ASR system will provide the correct answer is maximal. But each selection error, even a small one, will reduce this probability considerably. Now the question changes to: how to make a sufficiently correct selection? We might guess a word string and search for evidence of that word-string. When the words are correctly guessed it is likely that a sufficient fraction of the evidence can be selected. Unfortunately, there is an effectively infinite number of possible word strings, so the search space is huge; and it will be extremely difficult to determine what the best combination of word-string and selection is.

We are stuck in the speech-in-noise-paradox, or more general:

(1.5) The signal-in-noise-paradox: a correct recognition result is possible with a correct selection of evidence and a correct selection of evidence is possible with a correct recognition result.

Unfortunately, neither a selection, nor a recognition result is available. This paradoxical situation is a direct consequence of working with arbitrary input. The paradox represents a central problem of (speech) recognition, whether natural or artificial. Apparently, the human speech recognition system is very efficient in combining the correct bits-and-pieces of information. Building an artificial speech recognition system that solves the signal-in-noise-paradox, requires intimate knowledge of the functional units of this process and how they can be modeled computationally.

7. Although the signal-in-noise-paradox is relevant for natural systems, it has obviously been avoided. The paradoxical situation arises from the assumption that selection and recognition must be different processes (a direct consequence of the application of pattern-matching for speech recognition purposes). It is likely that our brain avoids the paradox by integrating selection and recognition.
1.5 Solving the Paradox

An obvious question is what it actually means that a sound source can be recognized or characterized. To recognize something entails that it can be classified as an instance of a certain class. Generally speaking:

(1.6) A sound source is *classifiable* as a member of a certain class if its signal has a sufficient fraction of the characteristic properties of that class, without being inconsistent with the class.

The set of characteristic properties of a signal like speech is defined by the laws of physics and by conventions between speakers. In the case of speech, meaning is a typical characteristic constraint. Furthermore:

(1.7) A *recognizable signal* has a set of properties that allows its classification. And the source that produced the signal is subject to characteristic (physical) constraints that are reflected in the characteristics of the sounds it can produce.

It is the task of a general sound recognition system to use these characteristics optimally.

Modern ASR systems map the input to the best matching template, but do not check if the input satisfies the characteristic constraints of the class of signals to be recognized. Such systems will produce nonsense if the user cannot ensure the correct type of input. To prevent this type of error an additional constraint must be introduced:

(1.8) An acceptable recognition result is a recognition result that satisfies the characteristic constraints/properties of its class.

The *best acceptable recognition result* is therefore the best recognition result that matches sufficiently well. Note that a recognition system can never be absolutely certain that a recognition result is correct. It can determine the best interpretation of the data and determine whether the recognition result is acceptable, but this is not a guarantee that the recognition result is correct. Yet

---

8. Note the small difference between recognition and classification. Both refer to the same process, but recognition is, per definition, impossible when a signal has not been presented before. In this context recognition and classification will be considered to refer to the final outcome of the system. A more general definition may allow recognition and classification to occur at all levels of processing.

9. Sound sources that are designed to be unclassifiable (on the basis of acoustic information alone) are hi-fi audio systems. These systems have hardly any noticeable physical constraints and can deceive listeners by producing any recognizable class of sound at will.
the fact that we are rarely confronted with far-reaching and negative consequences of an incorrect interpretation of speech, shows that this guarantee is rarely essential.\(^{10}\)

Being able to recognize or to characterize a sound source requires detecting it. In animals this has to happen as quickly as possible, since the sound source might require a rapid adaptation of behavior. The onset, here defined as the moment the difference between the absence and presence of the source is detectable, is therefore of great potential importance. The onset is by definition aperiodic. After the onset, the signal might be perceivable for some time. Since the signal is produced by a physical system, it cannot change more rapidly than the physics of the source (and the transmission medium) allow. This means that its signal changes in a continuous manner. Finally, at some moment the sound source will stop or become otherwise unnoticeable by the perceptive system. This produces another discontinuity: the offset. The total response of the perceptive system to a single sound source consists of an onset, often followed by a relatively slowly changing signal body and concluded by an offset. This leads to the definition of a sound event:

(1.9) A sound event is the audible physical signal that consists of the sequence of an onset, an optional continuous development, and an offset.

A sound event describes the input originating from a single source to the auditory system. Our auditory system usually receives multiple sound events at any given time. Prior to further analysis the auditory system is unaware of the existence of the individual sound events, it simply responds to the combined influence of all concurrent sound events. Yet, since the auditory system is in fact able to recognize multiple concurrent sound events, it apparently assumes that the input is produced by different sources that all produce classifiable signals. According to conclusion 1.7 this means that the auditory system assumes that all sound events show a characteristic set of properties that allow recognition. This leads to a basic assumption of a general sound recognition system:

(1.10) A general sound recognition system may assume that an arbitrary signal consists of a superposition of sound events of which each, in principle, can be recognized.

\(^{10}\) This forms further support for conjecture 1.2.
If the system is able to assign enough information from the sound events of a single source to a coherent representation that does not contain information of other sources,\textsuperscript{11} it can recognize the sound as if the other sounds were not interfering. \textit{This can solve the signal in noise paradox.}

(1.11) An \textit{auditory event} is the internal representation of information that can be estimated from a single sound event. Auditory events may not represent information of uncorrelated sources.

If auditory events can be estimated reliably, it is possible to reduce the task of a general sound processing system to a pattern recognition task through the set of combinations of auditory events. A number of combinations of auditory events might be acceptable in the sense of conclusion 1.8. Summarizing:

(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.

Conclusion 1.12 suggests that speech recognition requires much more than pattern matching:

(1.13) Speech recognition is an \textit{active} process that aims to find the most meaningful and most consistent interpretation of the signal.

Although this conclusion is used as a central design guideline for the innovations described in the work, it will not be addressed directly.

These conclusions are supported by experiments within the field of auditory scene analysis (see Bregman 1990 for an excellent overview). Bregman describes a large number of experiments in which acoustic evidence is either integrated into a single percept or segregated into multiple percepts. Bregman divides auditory scene analysis into \textit{primitive processes} and \textit{schema-based processes}. Primitive processes form a first processing stage. This stage functions independently of the types of sound sources, since it is based on simple and reliable physical cues in the signal. A typical cue for primitive processes is common onset: signal contributions that start at the same time are likely to belong to the same source. Another cue is based on the fact that frequency development of harmonics of a quasiperiodic source are correlated through time. Consequently, correctly coevolving harmonics tend to be

\textsuperscript{11} As will be shown in chapter 2 and section 4.7 it is generally impossible to form bottom-up representation \textit{without any} contamination from concurrent sound events. A noiseless reconstruction of the information of the sound event is only possible \textit{after} correct recognition.
Recognizing Arbitrary Sounds

integrated in a single percept. Schema-based processing forms a second stage. The schema’s represent our knowledge about different types of potentially meaningful sound. The activation of a schema requires the activity of primitive processes that comply to the regularities represented by the schema’s.

This work is mainly aimed at modelling important functional aspects of primitive auditory scene analysis. In particular this work is aimed at auditory event forming. Schema-based processing, although essential for robust speech perception, will be addressed (implicitly) in a proposal for robust speech recognition in section 7.2. The output of a limited implementation of a “primitive scene analysis stage”, as presented in chapter 2, will be evaluated with a traditional speech recognition system using HMM-based pattern matching. Section 6.1 provides the most elaborate form of primitive ASA in this work.

1.6 Design Basics

The numbered conclusions of the previous sections reduce the search space associated with the development of general sound recognition systems. Conjecture 1.1 is probably the most restrictive of all. If a recognition system has to function as often as possible, it is not allowed to make any unjustified prior assumptions that may prevent the system from reaching correct recognition results. Typical built-in assumptions of modern ASR systems that are often unjustified are (Junqua 1996):

- the input consists of a single source,
- the background noise is absent, stationary, or known,
- the training set matches the operating environment,
- the input can be recognized.

When all of these conditions are met, the system may work adequately, but when one of these conditions is violated the system will fail. Unfortunately, in most application environments, these conditions cannot be guaranteed without considerable effort. A general speech recognition system, of which the human auditory system is an example, remains almost unimpaired in these environments. Apparently, a general speech recognition system uses weaker prior assumptions than modern ASR systems.
(1.14) A recognition system that functions as often as possible must be based on a set of the weakest possible prior assumptions.

Applying the results of previous sections, one can propose a set of two basic assumptions:

1. all sound events show an onset, an optional continuous development and an offset (definition 1.9)
2. the (physical) constraints of the source that are reflected in the signal allow and determine recognizability (conclusions 1.7 and 1.10)

These basic assumptions form reliable starting-points for a recognition process. However, the ability to reach a correct recognition result is determined by the state of the perceptive system and the interaction between all concurrent sound events. Generally speaking:

(1.15) A recognition system is able to recognize a sequence of sound events of a single source correctly, if and only if it can assign a sufficient amount of informative evidence about the source to a single representation, i.e., if it can form a sufficiently informative set of auditory events corresponding to the sound event sequence of the source.

More evidence, i.e., greater redundancy, might facilitate processing, but will not change the recognition result. Less evidence, i.e., reduced redundancy, will eventually produce ambiguous results, incorrect results or no results at all.

The work of Fletcher, French, Steinberg and Galt (French 1947, Allen 1993) in the context of the articulation index showed that the local (in terms of time and frequency) signal-to-noise ratio is the sole determinant of whether or not reliable information can be derived from a speech signal. Reducing the global SNR will reduce the area of the time-frequency plane where reliable information can be derived from, up to a point where insufficient reliable information is left to guide a successful search.

**Limits of the measurement process**

Conform conclusion 1.15 the central problem of the recognition of arbitrary sounds is the assignment of the correct information to the correct representation. Physical measurement theory tells that any measurement includes an error. This error decreases with the inverse of the square-root of the number of measuring points. For stationary signals, the measurement error can be reduced to an arbitrary small value by increasing the duration or number of measurements. For nonstationary signals, like speech, this is
impossible since the signal changes between and/or during measurements. This means that any system that processes arbitrary, nonstationary signals has to deal with an unknown amount\textsuperscript{12} of noise on the values of all feature estimates. For signals of a single source this is usually not very problematic since most natural signals, including speech, are extremely redundant. But in the case of unknown signals, which is the case until the signals have been identified, it eventually becomes impossible to separate the measurement error from the effects of the interaction between sound events. This means that:

\begin{equation}
(1.16) \text{It is generally impossible to detect and estimate the values of individual features completely and/or sufficiently reliably.}
\end{equation}

This entails that individual features can never be trusted before an acceptable recognition result is reached. This is an extremely important operational problem for all measurement systems, including the mammalian sensory system.

The best solution is to hypothesize all likely features and feature values and work with the feature hypotheses as possible interpretations of the input. Although the reliability of individual feature hypotheses might be low, this is certainly not the case for combinations of feature hypotheses. A higher-level interpretation of multiple noisy features can be estimated with a considerably smaller error because it is based on more evidence (a larger context). As will be shown in this work, the pitch of a noisy signal can be hypothesized relatively reliably, although it might be difficult to find the individual harmonics on which the estimate is based and to estimate their frequency. Note that higher level interpretations are hypotheses as well.

In the case of pitch-estimation in noise, it is often impossible to determine the precise time of on- and offset and it is difficult to avoid octave errors. Multiple pitch-contour hypotheses, each differing in the on- and offset time and frequency range, might be necessary. As will be shown in the next chapter, each pitch-contour hypothesis leads to an auditory event hypothesis. Some combinations of auditory events stem from the same source, and may lead to acceptable recognition results (and possibly even the correct recognition result). Other sequences will be incomplete or might combine information of multiple sound sources. A good recognition system must, of course, be able to discard these sequences.

\textsuperscript{12} The amount of noise can only be estimated in the context of a (correct) recognition result.
To summarize the consequences of the physical limitations of the measurement process:

(1.17) The inability to estimate features of arbitrary signals reliably, forces a recognition system to work with hypotheses about features and feature-values.

(1.18) Higher level feature hypotheses are more reliable than the lower level hypotheses they are based upon and the whole is easier to recognize than the constituting features.\(^{13}\)

and conform conclusion 1.15:

(1.19) The first processing stages of the auditory system must result in a set of auditory events hypotheses that include the auditory events required to reach a correct recognition result. Later processing stage must select the best acceptable recognition result.

This thesis is aimed at the development of a number of techniques that, eventually, allow the formation of a set of auditory events hypotheses conform conclusion 1.19. A technique that approaches auditory event forming is presented in section 6.1.

### 1.7 Quasi-Stationarity and Continuity

This section addresses the justified application of quasi-stationarity and the importance of continuity to keep track of signal components of a single source.

(1.20) A signal component is a single physically coherent signal constituent that can be described by specifying the temporal development of frequency and energy (phase is optional). Signal components can be combined to form sound events.

Signal components refer usually to either harmonics, complexes of harmonics, or aperiodic contributions such as noise bursts and on- and offset transients (see section 4.4).

\(^{13}\) Note that conclusion 1.18 entails that the final disambiguation of noisy features is a top-down process.
Quasi-stationarity

Most speech signal processing techniques are based upon the quasi-stationarity assumption. This means that certain aspects of the signal, like amplitude and frequency content, can be modeled as originating from a process that is assumed to be constant over short periods\(^\text{14}\) (for speech a value of around 10 ms is usually chosen). The rationale for this assumption is that speech is produced by a physical system that cannot change infinitely fast (Young 1997). This is a perfectly reasonable assumption that is used extensively in this thesis. But the assumption holds exclusively for the signal of a single source (speaker). If a signal is produced by two speakers, it will change more rapidly and certainly differently than is allowed by the physics of a single vocal tract. Consequently, a form of quasi-stationarity that is only valid for a single vocal tract is not justified for mixtures of speakers and should be avoided. In an arbitrary, unknown environment, the situation is even worse, since signal contributions might exist for which quasi-stationarity is never a useful approximation. If quasi-stationarity is nevertheless applied, the induced approximation errors will degrade the combined signal irreparably and therefore reduce the probability to reach a correct recognition result.

Quasi-stationarity is often implemented by blocking the signal into frames and assuming (i.e., hoping) that the sequence of consecutive frames provides a sufficiently adequate description of the frequency-content of the target signal through time. Since the width of the frame (or the effective width of a window) is inversely proportional to the frequency resolution, a trade-off between temporal and frequency resolution is introduced. Signals in which frequency detail and temporal detail are both important cannot be processed optimally in a frame-based approach.

Another direct consequence of this approach is that the use of frames introduces discontinuities that make it difficult to determine the continuity (and consequently the existence) of underlying signal components. This in turn makes it more difficult to assign signal information of a single source to a single representation. The use of non-rectangular windows while discarding

\(^{14}\) The application of quasi-stationarity is very similar to the sample-and-hold process that is used to transform continuous signals into discrete signals. The correct application of quasi-stationarity is subject to the same conditions as the sampling process: the target features of the original signal must not contain frequencies that are represented ambiguously in the sampled version (Papoulis 1984, see The sampling theorem for stochastic processes). The dynamics of the source and the choice of features determine the sample frequency and thereby the maximal period of stationarity.
phase (the temporal information within the windowed signal) exacerbates this problem even more.\textsuperscript{15} To conclude:

(1.21) Quasi-stationarity, with a proper time-constant must only be applied to individual signal components or to complex signals, like the speech of a \textit{single} speaker, for which it holds. As long as the signal, or some selection of it, is not positively identified as suitable for the quasi-stationarity assumption, the application of quasi-stationarity is not justified and may lead to suboptimal or incorrect results.

This entails that virtually all speech signal processing techniques are ill-suited for use in a general acoustic environment. \textit{In particular, techniques like the Short Term Fourier Transform (STFT), Linear Prediction (LP) and frame-based filterbank methods should not be used on arbitrary signals.} These techniques are nevertheless applied to these signals, often without much success, or with successes on a very narrow range of applications (see Junqua 1996, which provides an overview of robust speech signal processing techniques). In the latter case, it is, again, the responsibility of the user to ensure that the input complies to the demands of the class of signals that can be dealt with.

Quasi-stationarity, with a proper time-constant, can only be applied safely to signal contributions of a single source. For an unknown mixture of sound events a more suitable form of signal processing is required.

**Continuity**

When aiming to develop a sound analysis system that assigns information of a single source to a single representation, one has to exploit the regularities of the source as well as possible. Unfortunately the regularities of the source are unknown, because the source is not yet classified. The system can, according to conclusion 1.14, only assume the weakest possible prior knowledge. But conform conclusion 1.9, it is safe to assume that any sound event has an onset, an optional continuous development and an offset. Consequently: all sound events that are not impulse-like show a continuously developing part.\textsuperscript{16} In the case of speech, most kinds of music and a wealth of natural signals, a continuous development is prominent most of the time. Only for some plosives like the /t/, the /k/ or the /p/ it might be argued that a continuous

\textsuperscript{15} More general: each irreversible operation destroys information and all irreversible operations should be avoided as long as the signal is not split into individual signal components.

\textsuperscript{16} It will be demonstrated in section 4.3 that even the response to an impulse can be characterized by a distinctive continuous development.
development is not visible in the signal. Utterances like *Why I owe you an hour?*, on the other hand, can be pronounced in such a way that the complete utterance forms a single continuous whole.

Continuity between on- and offset is a well-defined signal property that is shared by all sound events. Continuity, provided it can be justified from the signal, can therefore be exploited without any further knowledge of the type of signal. Continuity is therefore ideal to guide the assignment of acoustic evidence from individual sound events to auditory events.\(^{17}\) As long as a signal component shows a continuous development, it is likely to stem from a single source. This is a fairly safe conclusion, because the probability is small that uncorrelated sources give rise to signal components that extend each other smoothly. Furthermore, signal properties such as pitch-contours are continuous as well, and can help to group different signal components together: all harmonics of a single quasiperiodic sound event remain integer multiples of the fundamental frequency. Frequency contours consistent with a certain fundamental frequency contour are likely to belong to the same source, or, as is often the case with music, to multiple sources with a correlated temporal development.

Consequently:

\[
\text{(1.22) Continuity of signal components forms one of the most reliable cues for assigning information from a single source (or multiple correlated sources) to a single representation. While this process is not complete, continuity through time and frequency ought to be conserved.}
\]

In this light, it is to be expected that the first processing stages of the auditory system conserve continuity as well as possible. In the auditory system the transduction from sound, i.e., pressure fluctuations, to neural information is performed around a structure called the basilar membrane (see figure 2.2 or figure 3.1). The basilar membrane is a coherent physical structure that can be described by the physics of transmission lines. A transmission line is a structure that is continuous in both time and place, where in the case of the basilar membrane each position corresponds to a certain frequency that it responds best to. Consequently:

\[
\text{(1.23) The basilar membrane transduces acoustic vibrations to neural information and conserves continuity in time and frequency (via its correspondence to place) for further processing.}
\]

\(^{17}\) Although powerful, the use of continuity is in general neither required nor sufficient.
This property forms the basis of all new techniques described in this work.

Furthermore, each signal component tries to enforce its frequency and phase upon the basilar membrane. This succeeds best at the basilar membrane positions that are most sensitive to its frequency. Some signal components may entrain a region so efficiently that they are locally dominant.\textsuperscript{18} This entails that regions of the time-place plane where the local signal-to-noise ratio is favorable will be dominated by the target. This is in accordance with conclusion 1.15 and the experimental and theoretical work of Fletcher, French, Steinberg and Galt that showed that the local SNR (and not the spectrum!) determines the intelligibility of (nonsense) words. Consequently:

(1.24) The separation of a sound into the individual signal components starts at the basilar membrane where each signal component tries to dominate the corresponding BM-region.

This conclusion is (implicitly) implemented in Computational Auditory Scene Analysis (CASA) systems like that of Brown (1994) and Cooke. Since these systems are not based on representations that preserve continuity as well as possible, they cannot exploit the continuous nature of signal development optimally. Most computational auditory scene analysis systems (Rosenthal 1998) are based on varying combinations of neurophysiological plausibility and functional considerations. Yet, although Brown acknowledges the importance of weak assumptions, current CASA is not based on the most rigorous application of the weakest possible basic assumptions; consequently these endeavors are unlikely to lead to general recognition systems.

1.8 Task

The previous sections provided a starting point for sound processing that can help to develop recognition systems that are suitable for a maximally wide range of environments and tasks. As stated in section 1.4, the signal-in-noise-paradox represents a central problem of modern automatic speech sound recognition. It can be solved by assigning all estimable evidence of a single source to a single auditory event stream. When the combined evidence

\textsuperscript{18} Whenever the term local is used in the context of the time-frequency or time-place plane it will reflect locality in both time and frequency or both time and BM-position.
Recognizing Arbitrary Sounds

represents sufficient information, the signal can be recognized. This work focuses on a few essential steps in the auditory event estimation process. In particular:

- the identification of BM-regions where the influence of a single signal component can be estimated,
- the characterization of the information represented by these basilar membrane regions (typically the temporal development of energy and frequency) and
- the identification of combinations of signal contributions of the same source.

It is, according to conjecture 1.2, assumed that the most robust features of a speech signal represent the most important (linguistic) information. The work of Fletcher et al. showed that robust speech features correspond to regions of the time-frequency plane with a favorable SNR. Consequently: this work aims to identify and combine signal components of a single source that are likely to have a favorable (local) SNR. Therefore success is measured by:

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

When all three demands are satisfied (and conjecture 1.2 is valid) it is possible to develop speech recognition systems that are maximally robust because they are based on the most reliable (i.e., highest local SNR) and most informative features in the signal. Section 6.3 (figure 6.15 in particular) provides a quantification of a measure of success based on these three demands.

The main focus of this work is on the segregation of a signal in signal components. The integration of signal components can sometimes be justified from the signal, but often the combination of signal components depends on the application of linguistic or other knowledge which will be discussed in the form of a proposed recognition system in section 7.2. This work will only address the use of monaural information. Binaural processing (i.e., interaural time and level differences), although powerful, is not addressed.
1.9 Conclusions and Definitions

This chapter can be summarized by repeating the numbered conjectures, definitions, and conclusions:

(1.1) The human auditory system is optimized to function as often as possible in varying and uncontrollable acoustic environments.

(1.2) The most important (linguistic) information in a speech signal is represented by those signal properties that can be estimated by the human auditory system in the widest possible range of real-life conditions. These signal properties are, by definition, the most robust properties.

(1.3) An arbitrary sound is a sound of which no a priori knowledge is available: it can be any sound out of the set of all possible sound combinations.

(1.4) A general recognition system is a system that can recognize a signal that belongs to a certain target class, if and only if it is actually present in the input.

(1.5) The signal-in-noise-paradox: a correct recognition result is possible with a correct selection of evidence and a correct selection of evidence is possible with a correct recognition result.

(1.6) A sound source is classifiable as a member of a certain class if its signal has a sufficient fraction of the characteristic properties of that class, without being inconsistent with the class.

(1.7) A recognizable signal has a set of properties that allows its classification. And the source that produced the signal is subject to characteristic (physical) constraints that are reflected in the characteristics of the sounds it can produce.

(1.8) An acceptable recognition result is a recognition result that satisfies the characteristic constraints/properties of its class.

(1.9) A sound event is the audible physical signal that consists of the sequence of an onset, an optional continuous development, and an offset.

(1.10) A general sound recognition system may assume that an arbitrary signal consists of a superposition of sound events of which each, in principle, can be recognized.

(1.11) An auditory event is the internal representation of information that can be estimated from a single sound event. Auditory events may not represent information of uncorrelated sources.

(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.
Recognizing Arbitrary Sounds

(1.13) Speech recognition is an active process that aims to find the most meaningful and most consistent interpretation of the signal.

(1.14) A recognition system that functions as often as possible must be based on a set of the weakest possible prior assumptions.

(1.15) A recognition system is able to recognize a sequence of sound events of a single source correctly, if and only if it can assign a sufficient amount of informative evidence about the source to a single representation, i.e., if it can form a sufficiently informative set of auditory events corresponding to the sound event sequence of the source.

(1.16) It is generally impossible to detect and estimate the values of individual features completely and/or sufficiently reliably.

(1.17) The inability to estimate features of arbitrary signals reliably, forces a recognition system to work with hypotheses about features and feature-values.

(1.18) Higher level feature hypotheses are more reliable than the lower level hypotheses they are based upon and the whole is easier to recognize than the constituting features.

(1.19) The first processing stages of the auditory system must result in a set of auditory events hypotheses that include the auditory events required to reach a correct recognition result. Later processing stage must select the best acceptable recognition result.

(1.20) A signal component is a single physically coherent signal constituent that can be described by specifying the temporal development of frequency and energy (phase is optional). Signal components can be combined to form sound events.

(1.21) Quasi-stationarity, with a proper time-constant must only be applied to individual signal components or to complex signals, like the speech of a single speaker, for which it holds. As long as the signal, or some selection of it, is not positively identified as suitable for the quasi-stationarity assumption, the application of quasi-stationarity is not justified and may lead to suboptimal or incorrect results.

(1.22) Continuity of signal components forms one of the most reliable cues for assigning information from a single source (or multiple correlated sources) to a single representation. While this process is not complete, continuity through time and frequency ought to be conserved.

(1.23) The basilar membrane transduces acoustic vibrations to neural information and conserves continuity in time and frequency (via its correspondence to place) for further processing.

(1.24) The separation of a sound into the individual signal components starts at the basilar membrane where each signal component tries to dominate the corresponding BM-region.