Entropy and Speech

Mattias Nilsson

Sound and Image Processing Laboratory
School of Electrical Engineering
KTH (Royal Institute of Technology)

Stockholm 2006
Abstract

In this thesis, we study the representation of speech signals and the estimation of information-theoretical measures from observations containing features of the speech signal. The main body of the thesis consists of four research papers.

Paper A presents a compact representation of the speech signal that facilitates perfect reconstruction. The representation is constituted of models, model parameters, and signal coefficients. A difference compared to existing speech representations is that we seek a compact representation by adapting the models to maximally concentrate the energy of the signal coefficients according to a selected energy concentration criterion. The individual parts of the representation are closely related to speech signal properties such as spectral envelope, pitch, and voiced/unvoiced signal coefficients, beneficial for both speech coding and modification.

From the information-theoretical measure of entropy, performance limits in coding and classification can be derived. Papers B and C discuss the estimation of differential entropy. Paper B describes a method for estimation of the differential entropies in the case when the set of vector observations (from the representation) lie on a lower-dimensional surface (manifold) in the embedding space. In contrast to the method presented in Paper B, Paper C introduces a method where the manifold structures are destroyed by constraining the resolution of the observation space. This facilitates the estimation of bounds on classification error rates even when the manifolds are of varying dimensionality within the embedding space.

Finally, Paper D investigates the amount of shared information between spectral features of narrow-band (0.3-3.4 kHz) and high-band (3.4-8 kHz) speech. The results in Paper D indicate that the information shared between the high-band and the narrow-band is insufficient for high-quality wide-band speech coding (0.3-8 kHz) without transmission of extra information describing the high-band.

**Keywords**: speech representation, energy concentration, entropy estimation, manifolds.
List of Papers

The thesis is based on the following papers:


In addition to papers A-D, the following papers have also been produced in part by the author of the thesis:


Acknowledgements

I am now approaching the end of my Ph.D. studies, and I would like to take the opportunity to thank everyone who, in one way or the other, has supported me during this time. Although it is very likely that I forget someone, I still wish to mention a number of people who have been of great importance throughout this journey.

First of all, I would like to express my gratitude to my supervisor, Prof. Bastiaan Kleijn. Without his knowledge, clever ideas, and genuine interest in research, this thesis would not have been possible.

Next, I would like to thank all my present and former colleagues and guests at the Sound and Image Processing Lab: Anders Ekman, Andrei Jefremov, Arne Leijon, Barbara Resch, Christian Feldbauer, David Zhao, Davor Petrinovic, Dora Söderberg, Elisabet Molin, Ermin Kozica, Geoffrey Chan, Harald Pobloth, Kahye Song, Jan Plasberg, Jesús De vicente peña, Jonas Lindblom, Jonas Samuelsson, Karolina Smeds, Manohar Murthi, Martin Dahlquist, Moo-Young Kim, Peter Kabal, Peter Nordqvist, Renat Vafin, Shenghui Zhao, Søren Vang Andersen, Sriram Srinivasan, Volodya Grancharov, Yannis Agiomyrgiannakis, and Yusuke Hiwasaki. Your technical and social support cannot be overstated.

I especially would like to thank my previous and current office mates: Søren Vang Andersen, Manohar Murthi, Moo-Young Kim, and David Zhao for our numerous discussions on various topics such as information-theory, numbness, Kimchi, diapers and despair.

I would also like to devote a special thanks to Prof. Bastiaan Kleijn, Barbara Resch, David Zhao, Jan Plasberg, and Sriram Srinivasan for the proofreading of, and all the discussions on, the introduction and the papers included in this thesis.

Finally, using the highest number known by my eldest son, thirteen-eighteen-thousand-millions of kisses go to my beloved wife Kristina and my two wonderful boys Axel and Albin for their endless encouragement and unconditional love.

Mattias Nilsson
3 System description ............................................. A5
  3.1 Speech analysis ........................................ A5
  3.2 Speech synthesis ....................................... A13
4 Applications .................................................. A14
  4.1 Speech coding ........................................... A14
  4.2 Prosodic modifications ................................. A15
5 Experiments and results ................................... A16
  5.1 Implementation specific details ....................... A17
  5.2 Voiced and unvoiced separation ....................... A17
6 Concluding remarks ......................................... A21

References ..................................................... A21

B On the Estimation of Differential Entropy
from Data Located on Embedded Manifolds .............. B1
1 Introduction ................................................ B1
2 Entropy estimation ......................................... B3
  2.1 Entropy estimator based on quantization theory .... B4
  2.2 Estimation of the manifold dimension and entropy . B6
  2.3 Statistical properties .................................. B7
3 Experiments and results ................................... B10
  3.1 Artificial experiment I: Swiss roll .................... B10
  3.2 Artificial experiment II: ring torus .................. B11
  3.3 Real experiment: cepstral coefficients of speech ... B12
4 Conclusions .................................................. B15

References ..................................................... B26

C Intrinsic Dimensionality and its Implication
for Performance Prediction in Pattern Classification ...... C1
1 Introduction ................................................ C1
2 Problem formulation ....................................... C3
3 Feature space resolution .................................. C4
  3.1 Limiting resolution in plug-in estimators .......... C5
  3.2 Limiting resolution by noise addition ............... C5
4 Classification error probability estimation .......... C7
5 Experiments and results ................................... C9
  5.1 Varying intrinsic dimensionality .................... C10
  5.2 Pattern classification - artificial scenario ........ C11
  5.3 Pattern classification - speech vowels ............. C13
6 Concluding remarks ......................................... C15

References ..................................................... C15
D Gaussian Mixture Model based Mutual Information Estimation between Frequency Bands in Speech

1 Introduction ........................................ D1
2 Sound classes ...................................... D2
3 Modeling of the speech spectrum ............. D2
4 Information measures ............................... D4
5 Simulations .......................................... D6
   5.1 Simulation procedure ......................... D6
   5.2 Number of mixture components .......... D6
   5.3 Analysis of sound class sensitivity .... D7
6 Results ............................................. D7
7 Discussion and conclusions .................. D8
References ........................................... D9
### Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADPCM</td>
<td>Adaptive Differential Pulse-Code Modulation</td>
</tr>
<tr>
<td>APC</td>
<td>Adaptive Predictive Coding</td>
</tr>
<tr>
<td>ASR</td>
<td>Automatic Speech Recognition</td>
</tr>
<tr>
<td>AR</td>
<td>Auto-Regressive</td>
</tr>
<tr>
<td>CC</td>
<td>Cepstral Coefficient</td>
</tr>
<tr>
<td>CCR</td>
<td>Comparison Category Rating</td>
</tr>
<tr>
<td>CELP</td>
<td>Code-Excited Linear Prediction</td>
</tr>
<tr>
<td>DCT</td>
<td>Discrete Cosine Transform</td>
</tr>
<tr>
<td>DFT</td>
<td>Discrete Fourier Transform</td>
</tr>
<tr>
<td>DPCM</td>
<td>Differential Pulse-Code Modulation</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation Maximization</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian Mixture Model</td>
</tr>
<tr>
<td>i.i.d.</td>
<td>Independent and Identically Distributed</td>
</tr>
<tr>
<td>HNM</td>
<td>Harmonic plus Noise Model</td>
</tr>
<tr>
<td>KLT</td>
<td>Karhunen-Loève Transform</td>
</tr>
<tr>
<td>kbps</td>
<td>kilo bits per second</td>
</tr>
<tr>
<td>LER</td>
<td>Log Energy Ratio</td>
</tr>
<tr>
<td>LP</td>
<td>Linear Prediction</td>
</tr>
<tr>
<td>LPC</td>
<td>Linear Predictive Coding/Linear Prediction Coefficients</td>
</tr>
<tr>
<td>LSF</td>
<td>Line Spectral Frequency</td>
</tr>
<tr>
<td>MBE</td>
<td>MultiBand Excitation</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>MDCT</td>
<td>Modified Discrete Cosine Transform</td>
</tr>
<tr>
<td>MDL</td>
<td>Minimum Description Length</td>
</tr>
<tr>
<td>MFCC</td>
<td>Mel-Frequency Cepstral Coefficients</td>
</tr>
<tr>
<td>MI</td>
<td>Mutual Information</td>
</tr>
<tr>
<td>MLT</td>
<td>Modulated Lapped Transform</td>
</tr>
<tr>
<td>MNRU</td>
<td>Modulated Noise Reference Units</td>
</tr>
<tr>
<td>MST</td>
<td>Minimum Spanning Tree</td>
</tr>
<tr>
<td>NN</td>
<td>Nearest Neighbor</td>
</tr>
<tr>
<td>PCM</td>
<td>Pulse-Code Modulation</td>
</tr>
<tr>
<td>pdf</td>
<td>probability density function</td>
</tr>
<tr>
<td>pmf</td>
<td>probability mass function</td>
</tr>
<tr>
<td>PSOLA</td>
<td>Pitch-Synchronous OverLap-Add</td>
</tr>
<tr>
<td>RMS</td>
<td>Root Mean Squared</td>
</tr>
<tr>
<td>r.v.</td>
<td>Random Variable</td>
</tr>
<tr>
<td>SLB</td>
<td>Shannon Lower Bound</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal to Noise Ratio</td>
</tr>
<tr>
<td>ST</td>
<td>Steiner Tree</td>
</tr>
<tr>
<td>STC</td>
<td>Sinusoidal Transform Coding</td>
</tr>
<tr>
<td>STFT</td>
<td>Short-Time Fourier Transform</td>
</tr>
<tr>
<td>TTS</td>
<td>Text To Speech</td>
</tr>
<tr>
<td>TSP</td>
<td>Traveling Salesman Problem</td>
</tr>
<tr>
<td>VQ</td>
<td>Vector Quantization/Quantizer</td>
</tr>
<tr>
<td>WI</td>
<td>Waveform Interpolation</td>
</tr>
<tr>
<td>WSOLA</td>
<td>Waveform Similarity OverLap-Add</td>
</tr>
</tbody>
</table>
Part I

Introduction
Introduction

The novelist Charles Dickens has supposedly said that "Electric communication will never be a substitute for the face of someone who with their soul encourages another person to be brave and true."\(^1\) If he would have lived in the information technology oriented society of today he would perhaps have been a bit more careful with "will never be" in his statement. Equipment for video conferences has been available for some time and virtual reality communication systems simulating the complete audio-visual experience of face-to-face communication are emerging [76,86]. Due to the fast progress in technology it is very likely that we soon are as used to three-dimensional telepresence systems as we are now to existing communication systems such as the telephone, the television, the Internet, etc.

At a general level, the communication chain can be described by the following three consecutive blocks: a source, a channel, and a sink. The source produces a signal containing some information and the sink receives the parts of the information that have not been destroyed during the transmission of the signal through the channel.

Let us take the speech communication chain as an example. The speaker (source) formulates a linguistic message in his or her mind, and the brain maps it into neurological motor signals activating the muscles connected to the lungs, the vocal folds, and the vocal tract. This results in pressure changes in the vocal tract and at the lips, yielding the acoustic speech signal (source signal). Being a pressure wave, the acoustic signal propagates through the air and causes pressure changes at the ear drum of the listener (sink). In the ear, the relevant features of the signal are extracted and a neurological signal is generated and transported through the auditory nerve to the brain where the linguistic message is finally decoded [59,96]. In the previous example we considered that both the source and the sink were humans, but in speech communication applications such as automatic speech recognition (ASR) and text-to-speech (TTS) the sink and the source, respectively, are artificial.

This thesis is about the representation of source signals and the extrac-

\(^1\)www.whatquote.com/quotes/Charles-Dickens/11237-Electric-communication.htm
tion of information useful for both coding and recognition from the obtained representation. The focus is on speech, but we believe that many of the ideas and techniques presented can be applied to other sources.

The objectives of this thesis are to:

1. Develop a compact speech signal representation that can form a basis for speech coders and speech modification systems.

2. Develop tools and methodologies for measuring information useful for estimating performance limits in compression and classification given, e.g., a particular speech signal representation.

The remainder of this introduction is organized as follows. In Section 1 we discuss both general aspects of representing the speech signal and some particular existing speech signal representations. Section 2 focuses on the extraction of information from the signal representation. A summary of the contributions in this thesis is given in Section 3, which concludes the introduction.

1 Speech signal representation

In this section we discuss a generic approach of source signal representation and apply it to the speech signal. The section starts with providing a motivation for having a representation that captures the structure and characteristic features of the speech signal. We then discuss desired properties of a good representation, followed by a brief description of some existing speech signal representations. Finally, we present some useful tools that we utilize in Paper A to obtain an efficient representation of the speech signal.

1.1 Motivation

Let us start by defining what a representation is. In our context, a representation consists of a model, model parameters, and signal coefficients. The model is for instance an autoregressive model [96] of a particular prediction order, and the model parameters determine the pole locations of the autoregressive model. The signal coefficients of the representation characterize the mismatch, or equivalently the error, between the original signal and the signal as predicted from the model. It is possible to interpret the modeling of the speech signal as a mapping from the signal to the coefficients, where the mapping is determined by the model and the model parameters.

The main motivation for having a speech signal representation constituted by models, model parameters, and signal coefficients, is that it facilitates description of particular characteristics of the signal in a concise way. This makes it possible to develop more efficient systems for speech coding [45,67], speech modification (e.g., time- and pitch-scaling) [77,111,113],
and speech recognition [78] since tailor-made processing strategies can be
applied to the different speech signal features.
For instance, voiced (tonal) speech has a strong periodic structure due
to the vibration of the vocal folds. The vibration frequency of the vocal
folds is commonly referred to as the fundamental frequency or the pitch.
The periodic structure implies a high redundancy in the signal, which must
be exploited in efficient speech coding. Thus, modeling of the pitch of the
speech signal is valuable to coding. Representation of the pitch is also rel-
vent to speech modification applications such as time- and pitch-scaling.
Moreover, in speech coding, unvoiced (noise like) waveform descriptors can
be modeled by Gaussian noise (with properly matched gain and color) with-
out introducing any significant perceptual degradation to the reconstructed
speech signal [74]. The result indicates that unvoiced signal coefficients can
be compactly described since representing the (blocks of) coefficients by
their second-order statistics suffice. Similarly, in systems for text-to-speech
based on concatenative speech synthesis [112], identification of unvoiced
signal components facilitates time-stretching of unvoiced speech free from
artificial periodicity (i.e., avoiding periodic extension of unvoiced speech by,
e.g., phase randomization of the unvoiced signal coefficients).

1.2 Goodness of a representation

What is a good representation of a source signal? In general this depends
on the particular source signal and the type of application. However, it is
possible to define some generic properties that we would like a good source
signal representation to have, independently of the particular application.
Reasonable properties of the representation are that it is compact, complete,
and that it ”orthogonally” describes the relevant characteristics of the source
signal.

By a compact (or sparse) representation we imply that a signal can be
reconstructed at a low distortion from a representation that has relatively
few model parameters and few significant signal coefficients per time instant.
This is consistent with the energy of the coefficient space, after an energy-
preserving transformation, being concentrated in a small subspace, which
is advantageous for compression (cf. coding gain [49]).

Completeness of a representation means that the signal can be perfectly
reconstructed given the representation. Thus, transforms built from basis
(or frame) functions spanning the complete signal space are valuable tools to
achieve compact and complete representations. The concept of compactness
and completeness of a representation was previously discussed in [65,70]
in the context of speech and is consistent with the more formal objectives of,
e.g., minimum description length (MDL) principle [13], and rate-distortion
theory [30,108].

The third desired property of the representation is that the different
characteristics of the source signal are captured by almost independent parts of the representation. E.g., in the case of a speech signal, we prefer to have different parameters that describe the short-term dependencies due to the resonances of the vocal tract, the pitch-track (evolution of the pitch over time), and signal coefficients of the voiced and the unvoiced speech. This decomposition allows for high-quality manipulations of the speech signal. That is, if we alter the model parameters or signal coefficients, the output signal after synthesis should be a speech signal albeit a different one than the original.

1.3 Some existing speech representations

Speech coding research has been the primary driving force in finding more efficient representations of the speech signal. Comprehensive overviews on speech coding research can be found in, e.g., [45, 48, 50, 69, 104, 110]. A secondary driving force, of significantly less magnitude, is speech modification research, where high-quality time- and pitch-scaling is of great importance [113]. In the following, we briefly discuss some of the existing speech representations.

Historically, the speech coding methods can be split into two broad categories; waveform coders and voice coders (vocoders). The former category aims at replicating the speech waveform as accurately as possible, whereas the latter category synthesizes the speech signal using a parametric model of speech production. The properties of the waveform coders and the vocoders are in stark contrast to each other. For example, a waveform coder, such as the pulse code modulation (PCM) scheme [1,62], is capable of producing high-quality decoded speech. The expense is, however, that a high bit-rate (more than 16 kbps for speech sampled at 8 kHz) is required for the transmission of the encoded representation to achieve sufficiently high perceptual quality. The vocoders [44], on the other hand, facilitate speech coding at very low-bit rates (less than 4 kbps for speech sampled at 8 kHz), but the performance in terms of perceptual quality saturates at a relatively low level as the bit-rate of the vocoder increases.

Research on waveform coders and on vocoders has been motivated by making their corresponding representations compact and complete, respectively. In the case of the original PCM waveform coder, more compact representations are achieved by using the dependencies between adjacent speech sample amplitudes. Initially, prediction of the current speech sample amplitude was performed using the previous speech samples with fixed prediction coefficients. The actual encoding was performed using PCM on the prediction error. An example of such a coding scheme is differential PCM (DPCM) [50,102].

Since the dependencies between speech signal samples vary with both the vocal tract configuration (including lip radiation and glottal excitation)
and the speaker’s pitch, a fixed predictor cannot predict the speech signal samples efficiently at all times. Thus, to increase coding efficiency time-adaptive prediction coefficients were introduced (using the assumption that speech can be considered stationary during segments of 10-30 ms duration). The associated coding is referred to as adaptive predictive coding (APC), or more commonly, linear predictive coding (LPC) [2,12,102].

The vocoder was invented in 1939 by H. Dudley [35,104] and showed that speech can be represented compactly. Dudley's system generated synthetic speech by spectrally shaping a source signal that was either produced by a random noise generator or a pitch controlled oscillator. The spectral shaping simulated the vocal tract configuration and could be changed continuously through a resonance control unit. Although the original vocoder yielded a very compact representation of speech, the synthesized speech was far from natural-sounding. From the autoregressive (AR) modeling of the vocal tract (discussed later in Section 1.4), linear prediction (LP) follows naturally [80,81] motivating the LPC based vocoders [10]. To improve the quality of the synthesized sound, more elaborate models of the source excitation have been proposed.

The concept of modeling the voiced excitation by a sum of sinusoids was introduced in [55], and further refined into the approach sinusoidal transform coding (STC) [69,84]. In [7] the sinusoids were constrained by the pitch, and thus, the approach was referred to as harmonic coding. An interesting feature of the harmonic coding presented in [7] is that the error between the harmonic model and the original excitation is quantized and transmitted. Thus, in principle, the system in [7] forms a complete description of the speech signal. An alternative to STC that facilitates mixed voiced/unvoiced excitation is the multiband excitation (MBE) vocoder [25,54]. The MBE represents the excitation spectrum by a pitch (fundamental frequency), a voiced/unvoiced decision for each harmonic of the pitch frequency, and the phase of each voiced harmonic.

The need for improved compactness of the waveform coders and completeness of the vocoders as discussed above has led to hybrids of the two coder categories. The most well known hybrid coder is the code-excited linear prediction (CELP) scheme [11] that followed from the idea of the multipulse excitation [9]. The principle of CELP combines linear-predictive analysis-by-synthesis [105] with codebooks (vector quantizers) representing the excitation signal. CELP is the core of most of cellular codec standards of today [3–5,40].

Another coding scheme capable of producing high-quality decoded speech at very low bit rates is the waveform interpolation (WI) coder [66,67]. In WI, the one-dimensional speech signal is represented by a two-dimensional surface of so-called characteristic waveforms. Each characteristic waveform consists of speech segments of length one pitch period time-scaled to a fixed normalized length and periodically extended. A major
advantage of WI over CELP is that the two-dimensional surface of characteristic waveforms in WI facilitates efficient voiced-unvoiced decomposition which has shown beneficial for coding [74] due to the different perceptual characteristics of voiced and unvoiced sounds.

The discussion so far has been related to speech coding, but also the area of speech modification has influenced the research on speech representations. In speech modification, compactness of the representation is generally of less importance compared to speech coding. An exception is TTS systems based on concatenative speech synthesis, where the speech sound database has to be efficiently stored [112]. For speech modification it is more important that the different characteristics of the speech are described by almost independent parts of the representation since this facilitates flexible manipulation of the speech signal, as mentioned in Section 1.2. One such speech representation is the harmonic plus noise model (HNM) [77,111,113]. Similarly to the harmonic coding in [7], the speech signal is decomposed into a harmonic part and a noise part. The harmonic part of the signal is modeled by harmonically related sinusoids with linear time-varying complex amplitudes, and the noise part is modeled by modulated colored Gaussian noise (following the energy envelope of the original signal over time). Besides the modeling of the noise part, HNM differs from the harmonic coding approach by having pitch-synchronous analysis and synthesis.

1.4 Modeling tools

In this section we present some general modeling tools useful for representing speech. In addition to the AR-modeling, mentioned in the previous section, the tools we present are: constant pitch warping, transformation, and best-basis selection. Transformations and best-basis selections are important tools for obtaining compact signal representations. Further, the pitch is a key feature of the speech signal. The high signal redundancy it implies facilitates a compact speech representation if the information is properly used. The pitch varies continuously as a function of time, and the tool of constant pitch warping, time-scales (warp) the speech signal into a (warped) signal that has a constant pitch. This facilitates both pitch-synchronous processing and the usage of efficient (fixed) transforms, which is of importance to coding and modification of the speech signal. The section ends with a discussion on how to combine the modeling tools to have a compact and complete representation of the speech signal, as proposed in Paper A.

Autoregressive modeling

To capture the short-term dependencies between speech signal samples, the autoregressive model is an efficient tool commonly used in current speech
coding systems, including adaptive differential pulse code modulation (AD-PCM) [2,102] and CELP [3-5,11,40]. The AR model of a discrete-time signal \(s(n)\) can be expressed as [94]

\[
s(n) = \sum_{m=1}^{M} a_m s(n - m) + e(n),
\]

where \(M\) is the model order, \(\{a_m\}_{m=1}^{M}\) is the set of parameters specifying the AR model, and where \(e(n)\) denotes the (prediction) error term. The parameters \(\{a_m\}_{m=1}^{M}\) are often referred to as linear prediction coefficients, since they are commonly estimated through linear prediction analysis [80, 81,96]. The AR-model specifies a prediction of the current sample \(s(n)\) given a weighted sum of previous samples, and the optimal weights \(\{a_m\}_{m=1}^{M}\) are found by minimizing the energy (variance) of the prediction error. In the frequency domain, the AR-model results in an all-pole transfer function [96]. The power-spectrum of the all-pole transfer function describes the overall (smoothed) distribution of the signal energy along the frequency axis using few coefficients (thus a compact representation). Since the empirical short-term statistics of speech vary over time, the AR model parameters are commonly estimated from short speech segments (typically of 10-30 ms duration) using either the autocorrelation or the covariance LP analysis methods [96]. A historical perspective on LP analysis applied to speech can be found in [8].

Instead of viewing the AR modeling as merely a simple, yet efficient, model to capture dependencies between speech samples, it has been shown that the AR model can also be justified from a physical speech production perspective. Fant [43] did pioneering work in this area, and proposed the so-called source-filter decomposition of speech production [43] shown in Figure 1. The basic premise of the source-filter decomposition is that the speech signal is modeled by filtering an excitation source signal through vocal-tract and lip radiation filters. As seen from Figure 1, the excitation is a weighted sum of a pitch-pulse train (pulses spaced inversely proportional to the pitch and shaped by the glottal model) and random noise. The most well-known physical modeling of the vocal tract is the tube model [43]. A discrete-time model based on the concatenation of tubes of various diameters can be shown to be equivalent to an autoregressive (AR) model of the speech signal [32,43,96].

**Constant pitch warping**

In voiced speech the duration and shape of the pitch cycles generally change slowly, and the long-term dependencies are strong (high redundancy). Therefore, it is essential to take advantage of the pitch to get a compact representation of the speech signal. Since the pitch is continuously
varying, the design of transforms that maximally concentrate the energy of the signal coefficients is difficult. However, if the speech signal is warped into a signal of constant pitch, the design of these transforms is simplified. The parameters specifying the mapping between the original and warped signal domains are then a part of the speech signal representation.

To enable warping of the speech signal, we require access to an accurate description of a so-called warping function, relating the original time domain and the (warped) time domain in which the pitch is constant. In [98] the continuous-time warping function is modeled by cubic $B$-splines, and the optimal $B$-spline coefficients are assumed to be the ones that minimize the squared error between the speech signal and the speech signal delayed by one pitch period. Under the assumption that both the continuous speech signal and the warping function are well described with $B$-splines, a standard gradient descent method can be used to find an estimate of the warping function [63, 98]. Since, the warping function is related to the (instantaneous) pitch, standard pitch estimation methods [22, 68, 85] can be used to obtain an initial estimate of the warping function for the gradient descent method.

The actual warping can either be performed by methods of interpolation (resampling) in the time domain using the warping function as done in [98], or by zero-padding pitch cycles as done in [41]. The latter approach requires pitch marking such that the zero-padding can be performed within the closed phase of the glottal cycle. Moreover, if we want to apply a frequency transform to the warped signal, the zero-padding approach to warping is disadvantageous since it yields an oversampled frequency representation. This is not the case if we use an interpolation based warping approach.

---

Figure 1: Source-filter model of speech production.
Transforms and filter banks

A desired property of the representation of the speech signal is compactness. One aspect of compactness is that the energy of the signal coefficients, after an energy-preserving linear transformation, is concentrated into a few of the signal coefficients only. This generally facilitates a low distortion when reconstructing the signal from only a few significant signal coefficients. Transforms and filter banks can be used to achieve energy concentration.

Let us start the discussion on transforms by considering the Karhunen-Loève transform (KLT) which is commonly used in speech signal processing both for coding [51, 64] and enhancement [37, 87]. The KLT is a unitary transform\(^2\) \(U_{\text{KLT}}\) that when applied to a \(d\)-dimensional random variable \(X\) (representing e.g., a speech signal block) diagonalizes the covariance matrix of \(X\) (or equivalently the autocorrelation matrix in the case of \(X\) being zero mean). That is, if \(X\) is zero mean the autocorrelation matrix of the transformed vector \(Y = U_{\text{KLT}}X\) becomes

\[
E[YY^T] = U_{\text{KLT}} E[XX^T] U_{\text{KLT}}^H = \begin{bmatrix}
\lambda_1 & 0 & \cdots & 0 \\
0 & \lambda_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_d
\end{bmatrix},
\]

where the superscript \(^H\) denotes conjugate transpose. From (2) we note that the KLT is computed from an eigenvalue decomposition of the autocorrelation matrix \(E[XX^T]\) and that the rows of \(U_{\text{KLT}}\) consist of the eigenvectors of \(E[XX^T]\).

The concentration of energy obtained by a unitary transform is commonly evaluated using the coding gain defined as [49, 62]

\[
G_{\text{coding}} = d^{-1} \sum_{i=1}^{d} \sigma_{X_i}^2 \left(\prod_{i=1}^{d} \sigma_{Y_i}^2 \right)^{-1},
\]

where \(\sigma_{X_i}^2\) and \(\sigma_{Y_i}^2\) are the component variances of the \(d\)-dimensional random variable before and after the transform, respectively. It can be shown that the coding gain is optimized by the KLT [51], and thus, beneficial in transform coding.

As seen from (2), the KLT is signal dependent since it is constructed from the eigenvectors of the covariance/autocorrelation matrix. This is a disadvantage in the context of coding because the transform itself has to be transmitted. Interestingly, if we consider \(d\)-dimensional signal blocks from a stationary signal the KLT becomes the discrete Fourier transform (DFT) as the block length \(d\) tends to infinity [49] (cf. complex exponentials

---

\(^2\)A unitary transform \(U\) is a transform that satisfies \(U^H U = I\), where \(I\) denotes the identity matrix.
are eigenfunctions of linear time-invariant LTI systems [90]). This result justifies the commonly used approximation of the signal dependent KLT by the signal independent DFT or discrete cosine transform (DCT) [103].

The unitary transforms of the KLT, the DFT, and the DCT are constituted by a set of orthonormal basis functions. These transforms perform a one-to-one mapping between \(d\)-dimensional coefficient spaces. It is also possible to construct transforms that perform a one-to-one mapping between a \(d\)-dimensional space and a space of dimensionality higher than \(d\). The formalism that discusses this is the so-called frame theory [82].

Frames are a generalization of bases. Frames, and implicitly also bases, are efficient means for showing the perfect reconstruction property of uniform filter-banks (UFB). Let \(f\) denote a band-limited discrete signal (function) in the discrete Hilbert space, \(l^2(\mathbb{Z})\). A frame consists of a set of functions \(\{\gamma_k\}_{k \in \mathbb{K}}\) (\(\mathbb{K}\) being a countable index set) that satisfies the frame condition, e.g., [82]

\[
A \langle f, f \rangle \leq \sum_{k \in \mathbb{K}} |\langle f, \gamma_k \rangle|^2 \leq B \langle f, f \rangle, \tag{4}
\]

where \(\langle \cdot, \cdot \rangle\) denotes the inner product and the scalars \(A > 0, B < \infty\) are the so-called frame bounds (constants). We call a frame where \(A = B\) a tight frame. Furthermore, we define the frame operation (\(\mathcal{U}\) denotes the operator) on the signal \(f\) as

\[
\mathcal{U}f(k) = \langle f, \gamma_k \rangle. \tag{5}
\]

The frame condition (4) forms a necessary and sufficient condition for \(\mathcal{U}\) to be invertible with a bounded inverse. The dual operator of \(\mathcal{U}\) is its pseudo-inverse, i.e.,

\[
\mathcal{U}^t = (\mathcal{U}^H \mathcal{U})^{-1} \mathcal{U}^H, \tag{6}
\]

where the superscript \(H\) denotes conjugate transpose. For a tight frame, \((\mathcal{U}^H \mathcal{U})^{-1} \mathcal{U}^H = \frac{1}{\mathcal{X}} \mathcal{U}^H\), and that the frame is its own dual in this case, except for a constant factor.

Because of the conditions of invertibility imposed by the frame condition (4), frames can be used efficiently to show the perfect reconstruction property of the UFBs [23]. The analysis stage of such a filter bank corresponds to frames constructed by generating the set of frame functions from regular time-shifts of a finite set of functions, \(\{\gamma_{j,0}\}_{j \in \mathbb{J}}\). The transform coefficients are the outcomes of the inner products. Similarly, the synthesis stage can be seen as a summation, over all channels, of all dual (pseudo-inverse) frame functions for that channel; each frame function is scaled by its transform coefficient. Thus, the synthesis stage forms an expansion in dual frame functions. Let \(\gamma_{j,m}(n)\) denote a shifted frame function \(\gamma_j(n - m\mathcal{M})\) that is nonzero only over a finite time-support. Then
the (transform) coefficients of the frame expansion can be expressed by the inner product \( g_j(m) = \langle f, \gamma_j,m \rangle \), and the (perfectly) reconstructed signal becomes \( f(n) = \sum_j \sum_m g_j(m) \gamma^T_{j,m}(n) \), where \( \gamma^T_{j,m}(n) \) denotes a dual frame function with time offset \( m \) and frequency index \( j \). The analysis and synthesis stages of a uniform filter bank are shown in Figure 2.

![Figure 2: The analysis and synthesis stages of a uniform filter bank.](image)

There exist many different frequency transforms, e.g., DFT, DCT, short-time Fourier transform (STFT) [96], Gabor transform [15, 95] etc. A particularly attractive frequency transform is the modulated lapped transform (MLT) also referred to as modified DCT (MDCT) [36, 83]. A major reason for this is that the MLT facilitates the implementation of a critically sampled, i.e., renders as many output coefficients as input samples (or coefficients), perfect reconstruction uniform filter bank with well localized functions in time and frequency. This can for instance be achieved by combining smooth windows (that satisfy the power complementarity constraint\(^3\), needed for the perfect reconstruction) and DCT-IV functions [36]. That is, for this specific MLT the frame coefficients \( g_j(m) \) in Figure 2 can be expressed as the inner product between the function \( f \) and the frame function \( \phi_{jm} \), i.e.,

\[
 g_j(m) = \langle f, \phi_{jm} \rangle ,
\]

where

\[
 \phi_{jm}(n) = w(n - mM) \cdot \sqrt{\frac{2}{M}} \cos \left(\frac{(2j + 1)(2n - (2m + 1)M + 1)\pi}{4M}\right) .
\]

\(^3\)For a window \( w(n) \) with nonzero support for \( n = 0, \ldots, M - 1 \), the periodic extension of squared windows has to be constant.
where $M$ denotes the size of the time shift. The window $w(n - mM)$ in (8) can for instance be the square-root of the Hann window (it satisfies the power complementarity constraint) defined as

$$w(n - mM) = \begin{cases} \sqrt{\frac{1}{2} \left( 1 - \cos \left( \frac{\pi(n - mM + \frac{1}{2})}{M} \right) \right)} & n \in \Omega_n \\ 0 & \text{otherwise} \end{cases},$$

where $\Omega_n = [mM, ..., (m + 2)M - 1]$.

**Best Basis selection**

In best-basis selection, a library of orthonormal bases are used in combination with a cost function to match the basis to a given signal. The library of orthonormal bases can be for instance constructed from DCTs of different time support. To facilitate efficient coding it is desirable to have a cost function that measures the energy concentration of coefficients expressed in the terms of the new basis. A large number of different energy concentration cost functions exist, e.g., coding gain (3) based or entropy based [27], and the best choice depends on the particular application.

If we consider the speech signal, the effect of the best-basis selection using a library of DCTs of various time-support in combination with an energy concentration criterion is an automatic time-segmentation of the signal. Steady speech vowels are assigned DCT basis functions of long time-support and transients are assigned basis functions of short time-support. Thus, best basis selection [27] forms a powerful tool when seeking a compact and complete speech signal representation. For computational reasons the best-basis selection often starts from some fixed segmentation (of some minimum segmentation length) and, in a tree-structured bottom-up fashion, adjacent segments are merged if the new transform coefficients resulting from the merged segments yield an increased concentration in energy.

**An efficient tool combination for speech**

So-far, we have presented some useful tools for representing speech. In the following we briefly discuss the approach we have used in Paper A to combine these tools such that the resulting speech signal representation is compact and complete. The proposed arrangement of the modeling tools is shown in Figure 3, and consists of an initial LP analysis followed by constant pitch warping, a pitch-synchronous frequency transform, and a modulation transform. The modulation transform is a DCT transform applied to blocks of the time sequence of frequency coefficients from the pitch-synchronous frequency transform. The modulation transforms are in fact a best-basis selection where the block lengths are adaptively set to maximize some energy
concentration criterion. Besides the energy concentration potential of the modulation transform, it facilitates the identification of voiced and unvoiced signal coefficients beneficial for both coding [74] and prosodic modification [113]. This can be accomplished by assigning the coefficients of the low modulation bands to the voiced speech category. These coefficients represent the constant and slowly evolving components of the pitch-synchronous coefficients over time (block length of the modulation transform).

Figure 3: An efficient combination of modeling tools for representing speech.

The combination of the pitch-synchronous and the modulation transforms result in lapped frequency transforms [83], and similarly to all frequency transforms, they approximate the Karhunen-Loève transform (KLT) for stationary signal segments, as discussed previously. The KLT maximizes the coding gain [49,62] which can be seen as a particular energy concentration criterion. If the pitch is constant, the pitch-synchronous and the modulation transforms can be applied directly on the speech signal to achieve a highly energy concentrated representation. It is desirable to describe the variances of the signal coefficients after the transformations in an efficient way. Towards this goal we describe the spectral envelope by a parametric model. As is common in speech processing, we use the conventional AR model for this purpose. In practice, the pitch is not constant, but varying over time. However, by warping the speech signal of varying pitch into a signal of constant pitch, fixed pitch-synchronous and modulation transforms can be used. To facilitate perfect reconstruction the warped signal has to be oversampled. Thus, an increased efficiency of the AR-modeling is obtained if the LP analysis is performed prior to the warper, motivating the system structure in Figure 3.
2 Information extraction

In the previous section we discussed a generic approach of source signal representation and specifically applied it to the speech signal. This section considers the extraction of information from the signal representation obtained from source models. In particular, we are interested in information that is useful for the estimation of bounds on compression and performance of classifiers. The concept of entropy is a well-accepted measure of information, and we present in the following a detailed discussion on the definition of entropy, related measures, and existing entropy estimators. In practice, high-dimensional vector observations often associated with real-world sources, such as speech autoregressive model parameters or image tangent vectors, seem to have an effective space that can be parameterized with much fewer dimensions than the original vector [71, 106, 114, 115]. This causes severe problems in entropy estimation. We discuss these problems together with possible solutions at the end of this section.

2.1 Entropy

Entropy is Greek for "in transformation", and the term was originally introduced by Clausius in 1865 as a useful quantity for the definition of the second-law of thermodynamics\textsuperscript{4}. Clausius’ work on thermodynamics was followed by contributions from both Boltzmann and Gibbs, who during the 1870s established the connection between statistical mechanics and the thermodynamic entropy. In statical mechanics, entropy is seen as a macroscopic entity that quantifies the average disorder in an isolated system at a microscopic level.

To be more concrete, consider an isolated system with a fixed total energy and a fixed total number of particles \(N\). Each particle of the system can be in one out of \(M\) discrete energy states \(m \in \{1, 2, \ldots, M\}\), and the particles can exchange energy without loss. A microstate defines in this context a specific microscopic configuration of the system, i.e., one particular configuration of the \(N\) particles and \(M\) discrete energy states. Then Boltzmann showed that the entropy \(S\) is proportional to the logarithm of the number of possible microscopic configurations \(W\), under the assumption that all possible microstates of the system are equally likely and that constraints on energy mentioned above are satisfied, i.e.,

\[
S = k_B \ln(W),
\]

where \(k_B \approx 1.3806505(24) \cdot 10^{-23}\) denotes the Boltzmann constant, and where the number of possible configurations \(W = N!/(\prod_{m=1}^{M} N_m!))\), with

\textsuperscript{4}Entropy in an isolated system cannot decrease.
$N_m$ representing the number of particles that are in a specific energy state $m$.

A particular energy constraint allows a set of different distributions \{\(N_m\)\}. Under the fore-mentioned assumption that all microstates are equally likely and a given overall energy for all particles, one distribution is the most likely distribution. That is the distribution that corresponds to the largest number of microstates \(W\). This distribution is the so-called maximum-entropy distribution. For large systems (many particles in our example) the maximum-entropy distribution can be assumed to be the actual distribution (to be used for estimating macroscopic observables).

Applying Stirling’s approximation \(\log(Q!) \approx Q \log(Q) - Q\), which is valid for large \(Q\), to (10) the entropy can be expressed as

\[
S = -k_B N \sum_{m=1}^{M} p_m \ln(p_m),
\]

where \(p_m = N_m/N\) denotes the empirical probability that any particle is in a particular energy state \(m\).

### Information entropy

In 1948 Shannon published his pioneering work relating the uncertainty of messages to their limit of compression [107]. As a measure of uncertainty Shannon adopted the concept of entropy from statistical mechanics and developed the foundation of what is known today as information theory. In contrast to statistical mechanics, Shannon postulates a distribution (and not a uniform distribution of the microstates). Herein we define the (information) entropy and the differential entropy for a discrete and a continuous random variable, respectively.

Let \(\Xi\) be a discrete random variable (r.v.) that can take any outcome from a countable set of outcomes \(\mathcal{A}\). The set of outcomes is often referred to as the alphabet. Associated to the r.v. \(\Xi\) is a probability mass function (pmf) \(p_{\Xi}(\xi) \equiv Pr(\Xi = \xi)\) where \(\xi \in \mathcal{A}\). The information entropy of \(\Xi\) is then defined (in bits) as [107]

\[
H(\Xi) = -\sum_{\xi \in \mathcal{A}} p_{\Xi}(\xi) \log_2(p_{\Xi}(\xi)).
\]

In many cases the source under consideration, e.g., a speech signal, is analog in nature. The entropy of such a source as defined in (12) equals infinity, i.e., an infinite amount of bits is required for lossless encoding of the continuous-alphabet (analog) random variable. Interestingly, using the notion of a uniform quantizer (i.e., a quantizer where all quantization cells are identical) we can split the entropy of the indices from the quantizer.
into one term describing the intrinsic properties of the continuous random variable, and one term directly related to the cell volume of the quantizer. That is, for sufficiently small quantization cells the pmf of the indices can be approximated as $p_{\Xi}(\xi) \approx f_X(\xi) \Delta$, where $f_X(\xi)$ is the probability density function (pdf) of the continuous variable $X$ at position $\xi$, and $\Delta$ denotes the volume of the quantization cell. This yields

$$H(\Xi) = - \sum_{\xi \in \mathcal{A}} p_{\Xi}(\xi) \log_2(p_{\Xi}(\xi))$$

$$\approx - \sum_{\xi \in \mathcal{A}} f_X(\xi) \Delta \log_2(f_X(\xi)\Delta)$$

$$\approx - \sum_{\xi \in \mathcal{A}} f_X(\xi) \log_2(f_X(\xi))\Delta - \log_2(\Delta)$$

$$\approx - \int_{\Omega_x} f_X(x) \log_2(f_X(x)) \, dx - \log_2(\Delta)$$

$$\text{intrinsic property of } X$$

$$\text{effect of scale}$$

$$= h(X) - \log_2(\Delta), \quad (13)$$

where the entity $h(X) = - \int_{\Omega_x} f_X(x) \log_2(f_X(x)) \, dx$ is commonly referred to as the differential entropy of $X$. Note that the differential entropy can be negative, which contrasts with the entropy of a discrete random variable.

**Entropy based information measures**

For later purposes it is useful to define some other information measures related to differential entropy. Given the joint pdf of two r.v. $X$ and $Y$, $f_{XY}(x, y)$, the conditional differential entropy of $X$ given $Y$, $h(X|Y)$ is defined (in bits) as [30]

$$h(X|Y) = \int_{\Omega_x} \int_{\Omega_y} f_{XY}(x, y) \log_2 \left( f_X(x|Y) \right) \, dy \, dx, \quad (14)$$

where $f_X(x|Y) = f_{XY}(x, y) / f_Y(y)$ denotes the conditional distribution of $X$ given $Y = y$. The conditional differential entropy of (14) represents the remaining uncertainty of $X$ when $Y$ is given, averaged over the alphabet of $Y$.

By subtracting the conditional entropy $h(X|Y)$ from the differential entropy $h(X)$, we can quantify the information that $Y$ provides about $X$, or equivalently, the mutual information between $X$ and $Y$, i.e.,

$$I(X;Y) = h(X) - h(X|Y). \quad (15)$$

Using the joint and marginal pdfs of $X$ and $Y$, the mutual information can
be also expressed as

$$I(X;Y) = \int_{\Omega_x} \int_{\Omega_y} f_{XY}(x,y) \log_2 \left( \frac{f_{XY}(x,y)}{f_X(x)f_Y(y)} \right) dy dx. \quad (16)$$

From (16) we note that the mutual information is a symmetric measure in the sense that the information that $Y$ provides about $X$ is the same as the information that $X$ provides about $Y$.

**Shannon versus Renyi entropy**

An alternative to, or generalization of, Shannon’s definition of entropy is the Renyi entropy [97]. The Renyi entropy (of order $\alpha$) of the r.v. $X$ with pdf $f_X(x)$ is defined as

$$h_\alpha(X) = \frac{1}{1-\alpha} \log \left( \int_{\Omega_X} f_X^\alpha(x) \, dx \right), \quad (17)$$

where $f_X^\alpha(x)$ denotes the pdf of $X$ raised to the power $\alpha$. The Renyi entropy is sometimes advantageous due to mathematical tractability. For instance, Renyi entropy in combination with Parzen windows facilitates the derivation of gradient type algorithms in applications where entropy or mutual information is used as cost functions, e.g., blind deconvolution [38] and blind source separation [58]. Moreover, the graph-length of e.g., minimum spanning trees has shown to be closely related to the Renyi entropy [56], which has found applications in image registration [79,88].

As $\alpha$ approaches one, the Renyi entropy converges to the Shannon differential entropy $h(X)$, i.e.,

$$\lim_{\alpha \to 1} h_\alpha(X) = \frac{1}{1-\alpha} \int_{\Omega_X} f_X^\alpha(x) \, dx \left(\frac{1}{1-\alpha} \int_{\Omega_X} f_X^\alpha(x) \, dx \right)$$

$$= \lim_{\alpha \to 1} \frac{1}{1-\alpha} \frac{d}{d\alpha} \int_{\Omega_X} f_X^\alpha(x) \, dx \left(\frac{1}{1-\alpha} \int_{\Omega_X} f_X^\alpha(x) \, dx \right)$$

$$= \lim_{\alpha \to 1} - \frac{1}{f_{\Omega_X}^\alpha(x)} \int_{\Omega_X} f_X^\alpha(x) \log(f_X(x)) \, dx$$

$$= - \int_{\Omega_X} f_X(x) \log(f_X(x)) \, dx = h(X), \quad (18)$$

where we have used l’Hospital’s rule and assumed the validity of interchanging the orders of differentiation and integration.

In [107] Shannon stated three properties that are reasonable to require from a measure of uncertainty for a given pmf. First the entropy should be a continuous function of the pmf, i.e., an arbitrarily small change in
the pmf should not result in a jump in the entropy (cf. [101] for a formal definition of continuous functions). Second, if the pmf is uniform, then the entropy should be a monotonically increasing function of the cardinality of the discrete random variable. This is intuitively reasonable since with equally likely events the uncertainty about a particular event grows with the increasing number of possible events. Finally, the third desired property, according to Shannon, is that if a choice of an event can be split into two consecutive choices, the original entropy should remain the same and it should be possible to express it as a weighted sum of individual entropies. This is illustrated by the two example configurations in Figure 4, where we consider the uncertainty of the joint vector of two discrete r.v. $\Xi$ and $\Upsilon$. $\Xi$ and $\Upsilon$ have the possible outcomes zero and one, and thus, the joint vector $[\Xi, \Upsilon]$ has four possible outcomes. The configuration to the right

![Figure 4: Illustration of the desired conditioning property of the uncertainty measure according to Shannon.](image)

in Figure 4 is constructed by representing the four different events as a sequence of binary choices. If we let $\tilde{H}$ denote some measure of uncertainty, then for this example, Shannon’s third desired property implies that

$$\tilde{H}(\Xi, \Upsilon) = \tilde{H}(\Xi) + \sum_{\xi = \{0, 1\}} p_\xi(\xi) \tilde{H}(\Upsilon|\Xi = \xi) = \tilde{H}(\Xi) + \tilde{H}(\Upsilon|\Xi).$$

(19)

Shannon showed in [107, Appendix 2] that the only possible expression of the uncertainty that satisfies all three properties mentioned above is the expression of (12). That is, for the relation of (19) to hold $\tilde{H}(\Xi, \Upsilon) = H(\Xi, \Upsilon) = -\sum p_{\Xi\Upsilon}(\xi, \upsilon) \log_2(p_{\Xi\Upsilon}(\xi, \upsilon))$. Thus, the third property is not satisfied for the general Renyi entropy unless $\alpha = 1$ in (17), i.e., unless the Renyi entropy coincides with the Shannon differential entropy.
According to the source coding theorem [30], the entropy of an independent and identically distributed (i.i.d.) source is the lower bound on the average number of bits per source symbol needed for compression of the source without loss. This suggests that the relation in (19) is a desirable property since it facilitates the derivation of compression bounds including side information. Thus, the Shannon entropy clearly has an advantage over the Renyi entropy (for the case when $\alpha \neq 1$) in the field of coding theory.

### 2.2 Relation to coding and classification

A major motivation for estimating entropies and differential entropies is that they can be used to form bounds on the performance of source coders and pattern classifiers. Entropy and mutual information can also be used as cost functions in various signal processing applications, e.g., image registration or blind source separation as previously mentioned, and for optimal feature selection in speech recognition. In the following we discuss the role of entropy in both coding and classification.

#### Source coding

Source coding refers to the compression of signals. Source coders can either be of lossless or lossy compression type. As the name indicates, lossless compression implies that the source signal can be compressed and decompressed without introducing any distortion. Examples of commonly used techniques for lossless compression are the Huffman-coding [60] and the arithmetic-coding [99] schemes. Lossless coding using a finite number of bits is only possible when the source signal to encode is a sequence of discrete-alphabet variables.

When the receiver tolerates distortion to the processed signal, lossy compression techniques such as, e.g., scalar or vector quantization [53] can be used. A comprehensive overview on lossy source coding is given in [19]. Irrespective of the type of compression, i.e., lossless or lossy, it is always of interest to know/estimate the lowest average rate per symbol possible (or bounds on this rate) when designing source coders.

In the encoding of a sequence of discrete-alphabet source variables we assign to each source variable $\xi$ a codeword of length $l(\xi)$. Typically, we are only interested in codes that facilitate a decoding of a sequence of concatenated codewords. This is commonly referred to as the code being uniquely decodable. The unique decidability of the code places constraints on the set of possible codeword lengths $\{l(\xi)\}_{\xi \in \mathcal{A}}$. This constraint is formalized by the Kraft inequality: $\sum_{\xi \in \mathcal{A}} 2^{-l(\xi)} \leq 1$. Using the Kraft inequality, it is possible to show that the minimum average codeword length of a uniquely decodable code is bounded between the entropy of the source and the entropy of the source plus one bit. This is more formally stated in Theorem 1.
Thus, an estimate of the entropy provides useful knowledge about on how close to optimal performance the performance of a practical lossless coder is to.

**Theorem 1** The source coding theorem: The uniquely decodable code that minimizes the average codeword length, \( L = \sum_{\xi \in \mathcal{A}} p_{\Xi}(\xi)l(\xi) \), satisfies
\[
H(\Xi) \leq L = \sum_{\xi \in \mathcal{A}} p_{\Xi}(\xi)l(\xi) < H(\Xi) + 1. \tag{20}
\]

The theory specifying the optimal performance in lossy compression is referred to as rate-distortion or distortion-rate theory, and was also originally developed by Shannon in [108]. The rate-distortion function is a tight lower bound on the average bit rate \( R \) required to transmit a stationary process for a given average distortion \( D \). For a \( \mathbb{R}^d \)-valued r.v. \( X \) with pdf \( f_X(x) \) and a bounded distortion criterion \( d(x, \hat{x}) \), the rate distortion function is defined as
\[
R(D) = \inf_{\{f_{\hat{X}|X}(\hat{x}|x) : \mathbb{E}[d(X, \hat{X})] \leq D\}} I(X; \hat{X}), \tag{21}
\]
where \( f_{\hat{X}|X}(\hat{x}|x) \) specifies the statistical mapping between the original and the quantized variable, and where \( \mathbb{E}[d(X, \hat{X})] \) denotes the average distortion.

The rate-distortion function (21) is generally hard to determine analytically. An example of an exception is the combination of a univariate Gaussian source (zero mean and variance \( \sigma^2 \)) and a mean squared error distortion measure where the rate-distortion function is simply \( R(D) = 0.5 \log_2(\sigma^2/D) \). A numerical approach for the computation of the rate-distortion function is the Blahut algorithm [21]. Naturally, for a continuous alphabet source the Blahut algorithm yields an approximation of the rate-distortion.

An alternative approach that is usually simpler than finding the rate-distortion function is to determine a lower bound to it. The Shannon lower bound (SLB) is such a bound. The SLB can be defined for discrete-alphabet variables as well as for continuous-alphabet variables. For the latter case and a difference distortion criterion, i.e., \( d(x, \hat{x}) = d(x - \hat{x}) \), the SLB is defined as
\[
R_{\text{SLB}}(D) = h(X) - \sup_{\{f_{X-\hat{X}}(x-\hat{x}) : \mathbb{E}[d(X-\hat{X})] \leq D\}} h(X - \hat{X}). \tag{22}
\]
If the reconstruction variable \( \hat{X} \) is independent of the reconstruction error \( X - \hat{X} \) then the SLB coincides with the actual rate-distortion function and the SLB is said to be tight.
Classification

The aim of pattern recognition is to classify sensor data into a discrete set of predefined classes (patterns). Due to the curse of dimensionality [17], features extracted from the raw sensor data are commonly used for the classification task instead of the sensor data itself.

Through Fano’s inequality [42] it is possible to obtain a lower bound on the error probability of any classifier given the entropies of the classes conditioned on the features. Let the discrete r.v. \( Y \) represent the discrete classes we want to classify our sensor data into. Furthermore, let the continuous r.v. \( X \) represent the observations from which we classify \( Y \). Then the probability of error, \( P_e \), of a classifier is related to the conditional entropy \( H(Y|X) \) as [30]

\[
H_{P_e} + P_e \log_2 (|\Omega_Y| - 1) \geq H(Y|X),
\]

where \( H_{P_e} = -P_e \log_2 (P_e) - (1 - P_e) \log_2 (1 - P_e) \), and where \( |\Omega_Y| \) denotes the cardinality of \( Y \). Figure 5 shows an example of (23) for the case of two equally probable classes. The extreme points in Figure 5 are intuitive; if the class-conditional entropy is zero, then the lower bound on the classification error is zero. Conversely, if the class conditional entropy equals the class entropy the features do not provide any information and the error probability corresponds to chance only.

From (23) we note that selecting the set of features (represented by \( X \)) that minimizes \( H(Y|X) \) minimizes the lower bound on the classification error.
error probability. This is equivalent to seeking the features that maximize the mutual information (MI) between the features and the classes $I(Y; X)$. The MI $I(Y; X)$ represents the reduction (in bits) of the class entropy (uncertainty) $H(Y)$ that the features $X$ provide. Feature selection using the MI is treated in, e.g., [6, 14, 16, 24, 39, 116, 120].

2.3 Estimation of entropy

So-far, we have introduced the definition of entropy and how it relates to bounds on optimal coding and classification. The second main topic of this thesis concerns the actual estimation of the differential entropy. Existing techniques for (differential) entropy estimation can essentially be split into two categories; plug-in entropy estimators, and direct entropy estimators. The former category of estimators consists of a parametric or non-parametric estimate of the probability density function (pdf) followed by numerical integration, whereas the latter category utilizes the data directly to estimate the entropy avoiding the intermediate density estimation step. Examples of different pdf-estimation techniques used for the plug-in entropy estimates are histograms and kernels based models. In the category of direct entropy estimators we classify, for instance, the approaches based on the Euclidean distance between observations. A description of the different methods is given in this section.

Histogram based

An estimate of the pdf based on a histogram gives a piecewise constant approximation (it is constant over the histogram bin) of the true density. The resulting estimator of the differential entropy becomes (cf. the entropy to differential entropy relation in (13))

$$\hat{h}_{\text{hist}}(X) = \hat{H}(\Xi) + \log_2(V),$$  \hspace{1cm} (24)

where $\hat{H}(\Xi)$ is the estimator of the (quantization) index entropy of the histogram bins and $V$ is the volume of the histogram bins, assuming identical bins. The maximum-likelihood estimate of the probability of each bin $\hat{p}_\Xi(\xi)$ is equal to the ratio of the number of observations falling into the bin and the total number of observations. The corresponding estimate of the index entropy is $H(\Xi) = - \sum_{\xi \in \mathcal{X}} \hat{p}_\Xi(\xi) \log_2(\hat{p}_\Xi(\xi))$, where $\mathcal{X}$ denotes the set of indices. Since the number of histogram bins grows exponentially with the dimensionality, the practical use of the method is generally limited to the scalar or low-dimensional vector case. In [20, 46] the histogram bins are not of equal size and shape over the space, but instead they are individually selected as a function of the local structure of $X \in \mathbb{R}^d$, and the assumption of identical histogram bins in (24) (that resulted in the $\log_2(V)$ term) no longer holds and has to be compensated for.
Kernel based

The kernel-based pdf models used in plug-in entropy estimators can be either non-parametric or parametric. The Parzen windows approach is a commonly used non-parametric technique for density estimation [91]. In this case the pdf estimate is constructed by a superposition of windows centered around each observation of the random variable. Naturally, there are constraints on the windows to ensure that the pdf estimate is a valid pdf, that is it has to be non-negative and integrate to one. This can be guaranteed by letting the window itself be a density, e.g., a Gaussian. If we denote the window (or Kernel) centered around an observation \( x_n \) by \( \psi_N(x, x_n) \) the Parzen windows density estimate can be expressed as

\[
\hat{f}_{\text{Parzen},X}(x) = \frac{1}{N} \sum_{n=1}^{N} \psi_N(x; x_n).
\]  

The window \( \psi_N(x, x_n) \) is a function of the number of observations \( N \): the window width is selected to be inversely proportional to the number of observations. Figure 6 shows an example of Parzen-window based pdf estimation using Gaussian kernels.

![Figure 6: Example of pdf estimation using Parzen windows. The dashed curve shows the true pdf and the solid shows the Parzen window estimate using Gaussian kernels with standard deviation 0.1. The number of Parzen windows is 5 and 500 in the left and right plots, respectively. In the left plot the individual windows are shown (dotted curves).](image)

Plug-in entropy estimators require a multidimensional integration, which often is difficult or impossible to compute analytically. However, if we assume the random variable \( X \) to come from an ergodic source, we can replace the ensemble average \( h(X) = -E[\log_2 (f_X(X))] \) by the sample average
\[
\lim_{N \to \infty} -N^{-1} \sum_{i=1}^{N} \log_2(f_X(x_i)), \quad \text{and we obtain for the Parzen windows technique:}
\]
\[
\hat{h}_{\text{Parzen}}(X) = -\frac{1}{K} \sum_{k=1}^{K} \log_2 \left( \frac{1}{N} \sum_{n=1}^{N} \psi_N(x_k; x_n) \right),
\]
(26)
where \(K\) is the number of observations used for the stochastic integration and \(x_k\) denotes the \(k\)'th observation. The observations used for the stochastic integration above can be a subset of the true observations that was not used to form the density estimate, or artificially generated given the estimated pdf. In the former case, the estimator in (26) has a positive bias as \(K\) approaches infinity, since
\[
\lim_{K \to \infty} \hat{h}_{\text{Parzen}}(X) = h(X) + h(f_X \| \hat{f}_{\text{Parzen}}, X),
\]
(27)
where \(h(\cdot || \cdot)\) denotes the Kullback-Leibler distance [75] between the true and estimated distributions. Thus, by using a subset of the true observations in the stochastic integration, we have the advantage of knowing that our differential entropy estimate is an overestimate of the true differential entropy. Under some constraints on the smoothness of the window and the rate of decrease of the window width, convergence in mean square error of (25) can be shown [34, 91], and the bias term in (27) vanishes as \(N\) approaches infinity.

A commonly used parametric pdf model utilized in plug-in estimators of the differential entropy is the Gaussian mixture model (GMM). Similarly to the Parzen windows approach, the pdf is modeled by a weighted sum of kernels. However, in contrast to the Parzen windows, each kernel has an individual weight, mean, and covariance. Moreover, the number of kernels is an additional model parameter. Thus, the pdf using a GMM can be expressed as
\[
\hat{f}_{\text{GMM}, X}(x) = \sum_{l=1}^{L} \alpha_l \phi_l(x; m_l, C_l),
\]
(28)
where \(L\) denotes the number of components, \(\alpha_l\) is the component weight, and \(\phi_l(x; m_l, C_l)\) denotes the \(l\)'th Gaussian kernel with mean \(m_l\) and covariance \(C_l\). The parameters of the GMM, \(\{\alpha_l, m_l, C_l\}_{l=1}^{L}\), can be obtained using, for instance, the expectation maximization (EM) algorithm [33]. Figure 7 shows an example of pdf modeling using GMM. Assuming an ergodic source, we can form an estimator of the differential entropy using stochastic integration
\[
\hat{h}_{\text{GMM}}(X) = -\frac{1}{K} \sum_{k=1}^{K} \log_2 \left( \sum_{l=1}^{L} \alpha_l \phi_l(x_k; m_l, C_l) \right),
\]
(29)
where \(\{x_k\}_{k=1}^{K}\) is the set of observations not included in the estimation of the model parameters.
In contrast to the histogram approach, both the Parzen windows and the GMM based entropy estimators perform a smoothing of the observation space since the kernel widths are adapted to the number of observations and to the number of mixtures. In the case of Parzen windows the degree of smoothing is controlled by the kernel bandwidth and has to be manually set using, e.g., Silverman’s rule of thumb [109]. For the Gaussian univariate kernel the Silverman rule of thumb yields

$$\sigma_{\text{kernel}} = \frac{\sqrt{2}}{10} \left(0.9 \min (\hat{\sigma}_X, (q_{75} - q_{25})/1.34)\right)^{-1} N^\frac{1}{2},$$  \hspace{2cm} (30)

where $\sigma_{\text{kernel}}$ specify the standard deviation of the Gaussian kernel, $\hat{\sigma}_X$ denotes the estimate of the standard deviation of the data, $q_{25}$ and $q_{75}$ specifies the 25 percent respectively the 75 percent quantiles, and $N$ denotes the number of kernels.

**Euclidean distance based**

In [119] it was shown that an estimator of the differential entropy given a set of scalar observations can be derived from the distance between sorted observations (sorted from minus infinity to plus infinity). Intuitively this makes sense, since regions where the distances between the sorted observations are small are consequently also densely populated, and thus, closely related to the underlying pdf.

Sorting of observations is only possible in the scalar case. However, the Euclidean distance between observations in the vector space can be used for estimating the differential entropy [56,57,72]. In [56,57] it was shown that the graph-length obtained from, e.g., the minimum spanning tree (MST), the Steiner tree (ST), or the traveling salesman problem (TSP) (see [121]...
for more details on graph theory) is directly related to the Renyi entropy. Similarly, Kozachenko et al. [72] showed that the average nearest-neighbor distance relates to the Shannon differential entropy.

Let $\mathbf{x}^N = \{x_n\}_{n=1}^N$ define a set of $N$ $d$-dimensional independent identically distributed (i.i.d.) vectors in $\mathbb{R}^d$, and let an edge be the connection between two vectors of the set. Associated with each edge is a cost (in here the cost is the same as the Euclidean distance). A minimum spanning tree is defined as the set of edges that connects all vectors in $\mathbf{x}^N$ at minimum total cost. Thus, the length (or equivalently the cost), $L_{xN}$, of the tree is

$$L_{xN} = \min_{e \in \tau} \sum_{e} \|e\|^\gamma,$$

where $e$ denotes an edge in the tree $\tau$, $\| \cdot \|$ and $\gamma$ denote the Euclidean norm and power exponent, respectively. The most common algorithms for the minimum spanning tree search are the Kruskal [73] and Prim [93] algorithms. Figure 8 shows a minimum spanning tree of 512 random vectors containing the first and second mel-frequency cepstral coefficients MFCCs [31] of narrowband speech.

Figure 8: Minimum spanning tree of 512 random vectors containing the first and second mel-frequency cepstral coefficients of speech.

In [57] Hero and Michel exploit that the length of the minimum spanning tree can be related (for $d \geq 2$) to an estimate of the differential entropy ($\alpha$-
Renyi entropy):
\[
\hat{h}_\alpha(X) = \left(\log_2 \left( \frac{L_N^\alpha}{N^\alpha} \right) + \log_2 (\beta_{L,\gamma}) \right)/(1 - \alpha),
\] (32)

where the \(\alpha\)-entropy is controlled by varying \(\gamma\) as \(\alpha = (d - \gamma)/d\), and where \(\beta_{L,\gamma}\) is a density-independent constant only depending on the type of spanning tree (MST, ST, TSP, etc.), and \(\gamma\). A problem, however, is that \(\beta_{L,\gamma}\) is not known analytically, and has to be estimated from the data.

As previously mentioned, the Euclidean distance between nearest neighbors can be related to the Shannon differential entropy. In contrast to (32) the equivalent of the constant \(\beta_{L,\gamma}\) is known for the nearest-neighbor entropy estimator (cf. \(\beta_{L,\gamma}\) and the normalized moment of inertia for random codebooks [122]). The nearest-neighbor (NN) entropy estimator is defined as [72]
\[
\hat{h}_{NN}(X) = \frac{1}{N} \sum_{n=1}^{N} \log_2 (\|e_n\|^d (N - 1) \exp(C_{E}V_d)),
\] (33)

where \(\|e_n\|^d\) is the Euclidean distance between an observation \(x_n\) and its nearest-neighbor, \(d\) denotes the dimension, and \(V_d\) is the volume of a \(d\)-dimensional sphere with unit radius, i.e., \(V_d = \frac{\pi^{d/2}}{\Gamma(d/2)}\). The constant \(C_E \approx 0.5772\) in (33) is the Euler constant. Asymptotic properties of the NN entropy estimator have been investigated in [72,89,117,118].

### 2.4 Effects of manifolds for entropy estimation

A manifold is a mathematical space which is locally Euclidean, i.e., observing such a space from a close distance the space is approximately flat whereas when observed from further distance it may have a much more complicated structure. The surface of the Earth is an example of a two-dimensional manifold in a three-dimensional space. Locally, the distance between two neighboring Swedish cities such as Stockholm and Uppsala can be well approximated by the Euclidean distance in the three-dimensional space. However, if we would like to measure the distance between Stockholm and Sydney, Australia the Euclidean distance would be improper. Figure 9 shows two other two-dimensional manifolds embedded in the three-dimensional Euclidean space.

Manifolds have been observed in topics related to vision [100,106,114] and speech [71,115]. High-dimensional vectors of observations from these real-world processes have often been shown to be located on a lower-dimensional manifold embedded in the high-dimensional vector space. The dimensionality of the manifold is commonly referred to as the intrinsic dimensionality, and estimation methods for the intrinsic dimensionality are
Möbius strip  

Ring torus

Figure 9: Examples of two dimensional manifolds embedded in the three dimensional Euclidean space.

presented in e.g., [18, 26, 47, 61, 92]. This implies that the number of parameters needed to describe a point in the high-dimensional space is equal to the manifold dimension, assuming that the manifold structure is known. Dimension reduction techniques that consider an underlying manifold structure can be found in [100, 114].

The manifold structure of the data causes vanishing support of the probability density function, which in turn, when using the standard definition, results in an infinitely negative Shannon differential entropy. A more representative calculation of the differential entropy in this case is to first make a, possibly nonlinear, mapping of the data from the high-dimensional space onto a space of dimension equal to the intrinsic dimension of the manifold. The entropy calculation can then be performed in this lower dimensional space. The existence of manifold structures in the data has often been overlooked in entropy estimations, with the result that classical methods provide erroneous estimates of the entropy, since they assume the wrong intrinsic dimension (manifold dimension).

Recently, progress was made towards including the effect of data lying on a manifold in entropy estimation. Methods for estimating the Renyi entropy of data located on manifolds have been presented recently [28, 29]. The method in [28] is based on the construction of so-called geodesic minimal spanning trees obtained from pruning of the Isomap produced by the algorithm derived in [114]. In [29] the same authors reduce the complexity of their previous method by using $k$-nearest neighbor graphs instead of the geodesic minimal spanning trees. Contrasting with the work of [28, 29],
which provides estimation procedures for the Renyi entropy, our work provides direct estimation procedures for the Shannon differential entropy of data lying on a manifold. In Paper B, we start from high-rate quantization theory and arrive at a joint estimator of the intrinsic dimensionality and the Shannon differential entropy.

The assumption of a fixed manifold dimension over the embedding space is too strong in many cases. The manifold can be of varying dimensionality within the embedded space. In this case, the entropy estimation methods assuming a fixed dimensionality cannot be applied. The estimation of information theoretical measures in this context is addressed in Paper C.

It is important to emphasize that the manifold structures (of constant or varying dimensionality) of real-world processes are not likely surfaces in a strict sense because of the presence of noise (e.g., measurement or quantization noise). However, the noise can be so small that the practical behavior, given a finite set of observations, is very similar to the theoretical effect of infinitely negative differential entropy caused by manifolds.

3 Summary of contributions

In this thesis, we study the representation of speech signals and the estimation of entropy given observations containing features of the speech signal. In Paper A we present a compact representation of the speech signal that facilitates perfect reconstruction. The individual parts of the representation are closely related to speech signal properties such as spectral envelope, pitch, and voiced/unvoiced signal coefficients, which is beneficial for both speech coding and manipulation.

The canonical representation forms a natural basis for speech processing, including classification, bandwidth extension. When performing such processing, it is natural to wonder about any bounds on the performance of the processing. Such bounds are often based on information theory, and require estimates of the entropy. We address this problem in papers B and C.

The estimation of entropy is important since entropy is the core of the performance limit theorems in coding [52, 107, 108] and classification [42]. Paper B considers the estimation of the differential entropy in the case when the set of vector observations (from the representation), to be used for the estimation, lie on a lower-dimensional surface (manifold) in the embedding space. In contrast to the method presented in Paper B, Paper C introduces a method where the manifold structures are destroyed by constraining the resolution of the observation space in Paper C. This facilitates estimation of prediction error rates even when the manifolds are of varying dimensionality within the embedding space.

Finally, Paper D investigates the amount of shared information between
spectral features of narrow-band (bandwidth from 0.3 to 3.4 kHz) and high-band (bandwidth from 3.5 to 7 kHz) speech. The results in Paper D indicate that the information shared between the narrow- and high-band is insufficient to facilitate high-quality wide-band speech coding without transmission of some extra information describing the high-band.

Short summaries of the four research papers included in this thesis are presented below. Most of the derivations and all of the experiments in the following papers have been performed by the author of this thesis.

**Paper A: A Canonical Representation of Speech**

It is well known that usage of an appropriate representation of the speech signal improves the performance of speech coders, recognizers, and synthesizers. In this paper we present a representation of speech that has the efficiency, in terms of being compact, similar to that of parametric modeling, but additionally has the completeness property of signal expansions. The resulting canonical representation of speech is suited for a wide range of speech processing applications and we demonstrate this through experiments related to coding and prosodic modification.

**Paper B: On the Estimation of Differential Entropy from Data Located on Embedded Manifolds**

Estimation of the differential entropy from observations of a random variable is of great importance for a wide range of signal processing applications such as source coding, pattern recognition, hypothesis testing, and blind source separation. In this paper we present a method for estimation of the Shannon differential entropy that accounts for embedded manifolds. The method is based on high-rate quantization theory and forms an extension of the classical nearest-neighbor entropy estimator. The estimator is consistent in the mean square sense and an upper bound on the rate of convergence of the estimator is given. Because of the close connection between compression and Shannon entropy, the proposed method has an advantage over methods estimating the Renyi entropy. Through experiments on uniformly distributed data on known manifolds and real world speech data we show the accuracy and usefulness of our proposed method.

**Paper C: Intrinsic Dimensionality and its Implication for Performance Prediction in Pattern Classification**

Fano’s inequality relates entropy to the error probability of any classifier. It has been shown that feature observations extracted from real-world sources can have an intrinsic dimensionality different from the vector dimensionality. In this work, we show how this affects the estimation of entropy and consequently the prediction of classification error rates. We introduce a new
method that minimizes the effects due to the differences in dimensionality of the embedding space and the intrinsic dimension and demonstrate its performance.

**Paper D: Gaussian Mixture Model based Mutual Information Estimation between Frequency Bands in Speech**

In this paper, we investigate the dependency between the spectral envelopes of speech in disjoint frequency bands, one covering the telephone bandwidth from 0.3 kHz to 3.4 kHz and one covering the frequencies from 3.7 kHz to 8 kHz. The spectral envelopes are jointly modeled with a Gaussian mixture model based on mel-frequency cepstral coefficients and the log-energy-ratio of the disjoint frequency bands. Using this model, we quantify the dependency between bands through their mutual information and the perceived entropy of the high frequency band. Our results indicate that the mutual information is only a small fraction of the perceived entropy of the high band. This suggests that speech bandwidth extension should not rely only on mutual information between narrow- and high-band spectra. Rather, such methods need to make use of perceptual properties to ensure that the extended signal sounds pleasant.

**References**


[26] J. Bruske and G. Sommer. Intrinsic dimensionality estimation with opti-
mally topology preserving maps. *IEEE Trans. Pattern Analysis and Ma-


for monosyllabic word recognition in continuously spoken sentences. *IEEE


[33] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from

[34] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley,


[36] P. Duhamel, Y. Mahieux, and J. P. Petit. A fast algorithm for the im-
plementation of filterbanks based on "time domain aliasing cancellation".

[37] Y. Ephraim and H. L. Van Trees. A signal subspace approach for speech

Adaptive blind deconvolution of linear channels using Renyi’s entropy with


[40] European Telecommun. Standards Institute (ETSI). Enhanced Full Rate
EFR speech transcoding (GSM 06.60). 1996.

[41] G. Evangelista. Pitch-synchronous wavelet representation of speech and


