Objective Audio Quality Assessment Based on Spectro-Temporal Modulation Analysis

ZIYUAN GUO

Master’s Degree Project
Stockholm, Sweden 2011-11-25

XR-EE-SIP 2011:012
Abstract

Objective audio quality assessment is an interdisciplinary research area that incorporates audiology and machine learning. Although much work has been made on the machine learning aspect, the audiology aspect also deserves investigation.

This thesis proposes a non-intrusive audio quality assessment algorithm, which is based on an auditory model that simulates human auditory system. The auditory model is based on spectro-temporal modulation analysis of spectrogram, which has been proven to be effective in predicting the neural activities of human auditory cortex. The performance of an implementation of the algorithm shows the effectiveness of the spectro-temporal modulation analysis in audio quality assessment.
Acknowledgement

I would like to thank Professor Bastiaan Kleijn for the opportunity to do my thesis work in the Sound and Image Processing Lab at KTH and for advices regarding the work. I would also like to thank my supervisor Minyue Li for precious ideas and patient instructions and for the implementation of logistic regressor. Finally I would like to thank all the people at the lab for giving me support and inspiring environment.
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Human audio quality assessment process</td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>System diagram</td>
<td>10</td>
</tr>
<tr>
<td>2.1</td>
<td>Auditory model</td>
<td>13</td>
</tr>
<tr>
<td>2.2</td>
<td>Ear structure [32]</td>
<td>13</td>
</tr>
<tr>
<td>2.3</td>
<td>Basilar membrane [32]</td>
<td>15</td>
</tr>
<tr>
<td>2.4</td>
<td>Adaptation model with feedback circuits</td>
<td>17</td>
</tr>
<tr>
<td>3.1</td>
<td>Noise pattern</td>
<td>24</td>
</tr>
<tr>
<td>3.2</td>
<td>Burst pattern</td>
<td>25</td>
</tr>
<tr>
<td>3.3</td>
<td>Speech pattern</td>
<td>26</td>
</tr>
<tr>
<td>3.4</td>
<td>Modulation PSD contour of different signals</td>
<td>28</td>
</tr>
<tr>
<td>3.5</td>
<td>Modulation PSD contour of a speech in white Gaussian noise of different power</td>
<td>29</td>
</tr>
<tr>
<td>4.1</td>
<td>System structure</td>
<td>33</td>
</tr>
<tr>
<td>4.2</td>
<td>P.563 overall structure [19]</td>
<td>35</td>
</tr>
<tr>
<td>4.3</td>
<td>Diagram of CASP model in SIP’s system</td>
<td>36</td>
</tr>
</tbody>
</table>
List of Tables

1.1 Table of grades in the MOS scale .......................... 5
1.2 Table of grades in the DMOS scale ......................... 5
1.3 Table of grades in the CMOS scale ........................ 6

4.1 Test results .................................................. 34
## Contents

Introduction .................................................. 1

1 Audio Quality Assessment ............................... 3
   1.1 Audio Quality ........................................ 3
   1.2 Assessment ........................................... 4
      1.2.1 Subjective assessment .......................... 5
      1.2.2 Objective assessment ............................ 6
   1.3 Assessment System Structure ....................... 10

2 Human Auditory Perception ............................. 12
   2.1 Auditory System Overview ........................... 12
   2.2 Outer and Middle Ear ................................. 14
   2.3 Inner Ear ............................................ 14
      2.3.1 Cochlea .......................................... 14
      2.3.2 Basilar membrane ............................... 14
      2.3.3 Hair cell ....................................... 14
      2.3.4 Masking Effects ................................ 15
      2.3.5 Inner Ear Model ................................ 16
   2.4 Auditory Cortex Processing ......................... 17
      2.4.1 Per-utterance Cortex Processing ............... 18
      2.4.2 Per-frame Cortex Processing ................... 18

3 Machine Learning ........................................... 20
   3.1 Feature Extraction .................................. 20
      3.1.1 Per-frame extraction ............................ 21
      3.1.2 Per-utterance extraction ....................... 27
   3.2 Artificial Neural Network ........................... 27
   3.3 Scoring .............................................. 30

4 Evaluation .................................................. 32
   4.1 Test Setup .......................................... 32
   4.2 Results .............................................. 34

5 Conclusion .................................................. 38
Introduction

An important task of current telecommunication systems is to convey audio signals. In audio communication scenarios, signal processing tools such as compression and enhancement are commonly applied, and environmental impacts such as channel noise and delay are usually present. Therefore, the quality of the audio signals being transmitted is inevitably subject to degradation. Evaluation of audio quality is critical for the design and the operation of telecommunication systems. The traditional way of quality assessment is to collect peoples’ opinions on the quality in a well-controlled listening test. Such opinions are subjective and reflect true human perception. However, this subjective evaluation has limited applications, for example, when instant feedback is needed. So computer-based objective audio quality assessment is naturally desired. Many objective algorithms have been proposed and some of them have been standardized, for instance, ITU-T P.862 and ITU-R Rec. BS.1387.

Earlier development of objective audio quality assessment uses ad hoc approaches and lacks a systemic design. For example, some methods rely on the squared difference of two signals, and some others extract a vocal tract model from the input signal. More advanced objective audio quality assessment should exploit properties of human auditory system. In particular, many state-of-the-art systems uses general-purposed auditory models as the front-end, obtaining a so-called “internal representation”, on which they apply machine learning tools. The use of a delicate auditory model allows a quality assessment system to mimic the environment and a listener’s neural activities in a listening test.

The process of subjective audio quality assessment (Figure 1) consists of two parts: the auditory processing and the cognitive processing (cognitive mapping). Human auditory system performs the auditory processing in Figure 1, and the brain carries out the cognitive processing.

Both parts in Figure 1 can exploit existing technologies. Researches in psychoacoustics, physiology and neurology have led to many auditory models, which can be used in many auditory applications, e.g., automatic speech recognition
(ASR), audio compression, and audio quality assessment. For the cognitive processing, a closely related research area is machine learning.

Auditory model includes a peripheral part and may also includes a cortex part depending on the complexity of the model. The peripheral part is usually referred to as the ear model. The cortex part models auditory neurons in the cortex and auditory nerves that connect the ear and the brain. Periphery models have been extensively studied through anatomy, physiology and biochemistry. In contrast to the peripheral model, the procedure that happens in the auditory cortex and the form of “internal representation” of audio signals are less explored. The word “internal representation” denotes nerve excitation (see Figure 1) driven by signals that have passed through the human auditory system.

Recently, spectro-temporal modulation analysis is shown to be highly effective in serving as an “internal representation” of audio signals in human brain. Spectro-temporal modulation has psychologic and physical meanings. The has been used in many applications, such as intelligibility recognition [22], emotion recognition [34] and speech recognition [15].

Concerning the cognitive mapping in Figure 1, a machine learning scheme can be used to map the internal representation onto a quality score. As the internal representation usually is a large-sized matrix and has high dimensionality, feature extraction methods are essential before utility calculation that maps features onto quantities related to quality. An artificial neural network (ANN) is applied in our system to fulfill the utility calculation because ANN can perform almost all kinds of mapping. Finally, according to the quantities related to quality, scoring algorithm is responsible to predict the quality score by estimating its probability distribution.

The goal of this thesis is to implement and verify an audio quality assessment system consisting of an auditory model that utilizes spectro-temporal modulation and a machine learning scheme. We studied each functional unit and its mathematic model in an auditory model. Especially we implemented two spectro-temporal modulation analysis methods. One method is based on long-term power spectrum density. The other one is based on short-term DFT. In the machine learning scheme, we implemented and analyzed three feature extraction algorithms.

This thesis is organized as follows. Concepts about audio quality assessment are discussed in Chapter 1. A brief introduction of the proposed system is at the end. Chapter 2 introduces details of the auditory system structure and its mathematic model. In Chapter 3, three feature extraction algorithms are discussed and the machine learning scheme and scoring scheme are described. Test results under different schemes are shown and compared in Chapter 4.
Chapter 1

Audio Quality Assessment

The concept of audio quality attracts researchers’ attentions due to the wide spread of audio applications like recording and telecommunication. This concept is highly important in judging the performance of those audio applications. The evaluation of audio quality is traditionally carried out as a listening test. As the telecommunication applications has more and more requirements, such as the ability to assess the audio quality in real-time, a large number of objective measures have been proposed and implemented.

1.1 Audio Quality

The literal meaning of the term “audio quality” is about how good or bad audio signals are. The appearance of this term is largely due to the recording and the reproduction of sounds. From 1980, Compact Disc (CD) gradually replaced magnetizable tapes in the market. The format of audio CD, which is standardized by Sony and Philips, describes that each audio signal is sampled at 44.1 kHz and encoded by 16-bit PCM.

Both sampling and quantization cause information loss in real world due to the precision limit of all kinds of digital devices, so the reproduced signals are not the same as the original ones. Encoders and communication networks also cause changes between the input audio signals and the output audio signals. In the thesis, we concerned how to use objective algorithms to measure quality of audio signals that are processed by audio communication networks. Hence, audio quality in this thesis focuses on the changes between the unprocessed audio signals and the processed ones. There are cases that the unprocessed signals are not available, e.g., quality monitoring on one side of an end-to-end communication system. In these cases, people assess quality by comparing experience they learn from real life with what they heard.

At the early stage of the development of digital audio technology, quality degradation may mainly due to quantization errors and additive noises. There are several typical degradation factors in telecommunication systems and the audio codecs:

- Background noise. Background noise is a crucial factor that affects audio quality.
• Highly non-linear and signal-dependent distortion caused by audio encoding systems.

• Distortion caused by communication channel characteristics.

• Long delay and varying delay caused by heavy server loads or network access control.

• Non-sequential packets arrival caused by network congestion.

• Interruptions or clipping caused by bursts of packet loss.

• Inappropriate sidetone loss. Sidetone denotes the effect that sound picked up by the telephone’s mouthpiece is feedbacked to the headset receiver of the same telephone. If the sidetone loss is set too large, the receiver sounds are unnaturally silent, otherwise, the receiver sounds are too loud.

• Echo. It happens when the speaker’s speech signal returns with a sufficient delay in time so that it is perceivable from the normal sidetone.

• Crosstalk. It happens when the audio signal from one communication connection is coupled to another connection.

1.2 Assessment

The procedure of evaluating the audio quality is quality assessment. Evaluations are usually expressed as scoring on an ordinal scale for mathematical convenience. There are two categories of assessment: subjective and objective assessment. The audio quality should be evaluated by humans since the term “audio quality” is essentially subjective. Subjective assessment meets the subjective properties of quality and its results are considered truthful. Due to limitations of subjective assessment, researchers have developed objective assessment. Objective assessment usually tries to estimate the scores resulted from subjective assessment.

Audio quality is essentially multivariate. Some quality assessment methods are designed to measure one or more aspects of quality, which can be chosen from the degradation factors mentioned above or some intrinsic characteristics of sounds such as naturalness, coarseness, intelligibility and happiness. Since it is too complicated to identify all quality related factors and measure them, many quality assessment methods focus on an overall quality metric. In this thesis, we assumed that the overall quality is a summary of various aspects of a sound. We later refer to these aspects as “utilities”. The way of summarization, which is referred to as “scoring”, depends on the definition of the overall quality. An assessment method that judges too many aspects of quality becomes too complex to implement. Usually an assessment with a single metric is more commonly used than a multidimensional metric one. Although a single metric assessment method generally fails to give details of the quality, it is simple to implement and its judgement sufficiently represents listeners’ opinions. Therefore the system proposed in this thesis applies a single metric assessment method.
1.2.1 Subjective assessment

Subjective assessment is performed as a well-controlled listening test. A group of participants listen to audio materials according to a given metric. Generally, they give scores to represent the quality. It is noted that different persons may have different judgements due to some factors that are irrelevant to quality, such as the speaking language used in the test or the contents of the listening materials. Such irrelevant factors need to be removed or well controlled to decrease the variance in the quality rating. Another way to decrease the variance is to increase the number of participants. When the number of participants is sufficiently large, the total results can be seen as a representation of the true quality.

There are several typical subjective assessment tests. One popular and simple test is the Absolute Category Rating (ACR) test, which is a single metric test and standardized in ITU-T Rec. P.800. ACR tests are non-intrusive. After each piece of audio file is played, listeners should give their opinions about its quality using a scale shown in Table 1.1. The average of all the ratings is known as Mean Opinion Score (MOS).

| Excellent | 5 |
| Good      | 4 |
| Fair      | 3 |
| Poor      | 2 |
| Bad       | 1 |

A disadvantage of ACR tests is that its result may not have enough resolution due to the lack of reference signals. On the presence of a reference signal, Degradation Category Rating (DCR) is often used, which is also standardized in ITU-T Rec. P.800. In a DCR test, listeners are provided with a reference signal before they listen to each test signal. They need to rate audio materials using the scale presented in Table 1.2. The average of all the ratings is referred to as Degradation Mean Opinion Score (DMOS). This category of assessment is denoted as intrusive.

| Inaudible               | 5 |
| Audible, but not annoying | 4 |
| Slightly annoying       | 3 |
| Annoying                | 2 |
| Very annoying           | 1 |

In DCR tests, listeners always take the first signal as reference and evaluate the degraded quality of the second signal. An improved version of DCR is Comparison Category Rating (CCR) test. A CCR test is also a comparison test, in which the listeners similarly evaluate the quality of the second signal relative to the first one using a rating scale described in Table 1.3 and the average is called Comparison Mean Opinion Score (CMOS). However, the order that the reference and the test signals are presented to the listen is random.
A subjective test with more refined resolution is “MUlti Stimulus test with Hidden Reference and Anchor” (MUSHRA), which is recommended to assess the intermediate audio quality and is standardized in ITU-R Rec. BS.1534-1. In this test, a known reference and hidden anchors are included. Listeners listen to the signals in an arbitrary order and for arbitrary times, and rate the quality of test signals from 0 to 100.

1.2.2 Objective assessment

As the subjective methods are time-consuming, expensive and not suitable to be used in real-time applications, a lot of computer-aided objective methods are proposed and implemented. Since the results of subjective tests are considered to reflect the true quality on specific metrics, the aim of objective methods is to yield estimated qualities that are coherent with subjective test results, for example, MOS. Early objective assessment methods usually use ad hoc algorithms. For example, some methods try to exploit the correlation between audio test signals and quality scores by measuring the waveform difference. The problem with estimating quality using waveform comparison is that some changes give a large waveform difference but little audible distortion, for example, waveform inversion and phase shifting. More advanced objective assessment methods are built on well-designed auditory models that simulates auditory perception process.

Like the subjective assessment, the objective assessment has also two categories: intrusive and non-intrusive methods. Non-intrusive methods are more suitable for cases where reference signals are not available. Compared with intrusive methods, non-intrusive methods are relatively immature but have wider application situations. That is reason why this paper focuses the non-intrusive method. However, development of both categories has shed light on the same principle: the auditory model plays a critical role in an audio quality assessment system.

Development of Intrusive Models

Intrusive models have a processed signal and a reference signal as inputs and compare these two signals to derive scores according to specific quality metrics.

Although there are ad hoc approaches, perceptually motivated auditory models are dominant recently. Perceptually motivated auditory models are based on modeling the masking effect, which will be discussed in Section 2.3.4. Schroeder et al. [25] proposed a method that based on absolute masking thresh-
old to predict the audibility of coding noise. Further, Brandenburg [2] developed that method into the mean noise to masking ratio (NMR) by taking post-masking effect into account. However, NMR considers the differences between the reference and test signals in time domain as noise, which is not always true, for example, when the test signal is simply phase-shifted comparing to the reference. Karjalainen [13] proposed a more general model that based on auditory spectrum distance (ASD). This model compares the loudness of two signals in time-frequency domain. Compared with Schroeder’s approach, ASD changes the absolute masking threshold and considers temporal masking effect. Wang et al. [33] introduced a similar approach called Bark spectral distance (BSD). BSD computes the mean squared Euclidean distance on a Sone scale in the Bark bands. One drawback of BSD is that it does not take temporal masking effect into account. Beerends and Stemerding’s perceptual audio quality measure (PAQM) weights the difference in each time-frequency cell according to the power ratio of the reference and test signals in that cell. Perceptual speech quality measure (PSQM) adapts this approach. PESQ, which is standardized in ITU-T P.862.1, is the successor for PSQM described in ITU-T P.861. PSQM is basically used for evaluating the performance of speech codecs, while PESQ concerns not only codec distortion but also filtering, variable delay in the network and short-time distortion. So PESQ is suitable to assess the performance of 3.1 kHz handset telephony and narrow-band speech codec. The auditory perception model used in PESQ is systematically designed. It begins with an IRS filter that mimics the frequency response characteristics of the telephony receiver. Following the filter, a series of time alignment operations in both time and frequency domain are carried out to solve the variable delay between the reference signal and the test signals. The core of model is comprised of DFT, absolute hearing threshold, warping of frequency to Bark scale by summing the corresponding FFT bands and normalizing the summation, pitch power compensation for filtering and short-term gain variation, transform of loudness level according to Sone scale and cell-wise multiplication with asymmetry factors, which was originally proposed in PAQM.

Perceptual Evaluation of Audio Quality (PEAQ) is another standardized intrusive approach, which is used to judge small impairments in audio systems. PEAQ is described in ITU-R Rec. BS.1387 as an objective algorithm to characterize the results of a listening test that is standardized in ITU-R BS.1116. PEAQ basically adopts two psychoacoustic measurements. One is the masking threshold concept introduced in NMR. The other is the comparison of loudness as in ASD. The comparison mainly calculates differences on spectral bandwidth, harmonic structure, excitation envelope modulation and excitation magnitude modulation. PEAQ includes the model of the outer and middle ear. The time-frequency transform can be performed by FFT or filterbank. Frequency warping to Bark scale is also necessary. Masking effects are modeled as frequency domain spreading and time domain spreading. Internal noise, which simulates the noise in the nerves and the noise of blood flows, is included.

Besides approaches mentioned above, several methods process a spectrogram further with the aim of extracting modulation information. A widely used method is mel-frequency cepstral coefficients (MFCC), which is a cepstrum of a spectrogram. In speech processing, MFCC usually uses a short time cepstrum analysis:
1. Segment the original signal $x(t)$ into overlapping frames and produces $x_i(n), 0 \leq n < N - 1$, where $i$ represents the frame index and $N$ is the instant number of each frame.

2. Multiply each frame $x_i(n)$ with a window $w(n)$ and take the Fourier transform $X_i(f) = \mathcal{F}\{x_i(n) \times w(n)\}$. The energy spectrogram of $x(t)$ is defined as $S(i, f) = |X_i(f)|$.

3. Map the energy at each frequency bin in Hz scale onto a mel scale, $S(i, f) \rightarrow S'(i, m)$, where $m$ is the frequency in mel scale.

4. Apply a short time Fourier transform on the frequency axis, $MFCC(i, j) = |\mathcal{F}\{\log S'(i, m)\}|$.

As the cepstrum is the Fourier transform of the logarithm of the spectrogram of a signal, MFCC essentially exploits spectral modulation of the spectrogram and splits spectral components into linear sums of their cepstrum representations.

There are some other methods that utilize temporal modulation, for example, PEMO-Q [11]. PEMO-Q includes an auditory model, which was initially proposed by Dau et al. [6]. Dau’s model includes an auditory peripheral part and an auditory cortex part. In Dau’s model, a spectrogram is produced by passing a signal through a filterbank and the loudness level is adapted by a feedback loop. After envelope extraction and loudness level adaptation, each channel of a spectrogram is filtered by a temporal modulation filterbank. PEMO-Q regards the internal representation as a three-dimensional matrix that varies in time, frequency and modulation-frequency.

In summary, recent intrusive approaches utilize an auditory model that generally consists of time-frequency transform, frequency warping, loudness compression, masking threshold and modulation analysis.

Development of Non-intrusive Models

Intrusive models can achieve satisfactory results by comparing differences between reference signals and test signals. However, in many cases reference signals are not accessible, such as quality monitoring on one side of an end-to-end communication system. That is the reason why non-intrusive models are proposed and investigated.

There are several non-intrusive solutions. Gray [10] proposed a method based on a vocal tract model. He takes the assumption that most speech quality degradation, which are caused by speech processing systems or telecommunication networks, cannot be produced by the human vocal tract. This model tries to estimate the likelihood that a sound is produced by human’s voice production system. ITU-T P.563 extracts vocal tract parameters from a test speech file and other parameters such as interruption and additive noise.

Another tract of non-intrusive methods utilizes perceptual concepts. A method named perceptual linear prediction (PLP) was proposed by Liang and Kubichek [18]. This method considers three perceptual concepts: frequency warping, the equal-loudness curve and the intensity-loudness power law. Thus this method is consistent with human auditory perception. In this method, a fifth-order all-pole model is performed to approximate the spectrogram. PLP coefficients are transformed from the coefficients of the all-pole model and are
used in parametrization of a speech signal. The reference database is trained from PLP coefficients of clean speech signals. The time-averaged Euclidean distance between the PLP coefficients of a test signal and the nearest reference coefficients is calculated as an indication of degradation. Following the idea that is to measure the deviation of test speech from a statistical model, which is trained on clean speeches, Falk et al. used Gaussian mixture models (GMMs) to model the PLP feature vectors, which are obtained from clean speeches.

Another method, ANIQUE+ [14] was proposed by Kim. In contrast to PLP, ANIQUE+ delicately models both the periphery and cortex of the human auditory system. This approach processes input signals by an auditory model that includes a temporal modulation filterbank. After the auditory processing, it analyzes the power distribution of modulation spectrum. ANIQUE+ regards power in non-articulation modulation frequencies as distortion and the power distribution is considered as an indication that quantifies the degree of quality in speech signals.

Compared with the auditory model, there are several weaknesses of the vocal tract model. For example, it cannot evaluate perceptual effects, which are utilized by many encoders. Additionally, the assumption taken by Gray’s vocal tract model indicates that this method is only suitable for speech quality assessment and is not capable of assessing music quality. The majority of non-intrusive methods apply a perceptually motivated auditory model, which is consistent with human assessment process, and have achieved good and effective results. For example, the PLP method uses perceptual concepts and ANIQUE+ uses a complete auditory model. ANIQUE+ outperforms the PLP method for ANIQUE+ has a more advanced auditory model than the PLP method. Following the major tract, the non-intrusive approach in our system should adapt successful auditory models, which are utilized in intrusive approaches, to generate effective internal representations rather than focusing on the vocal tract model. In other words, an objective non-intrusive system on this tract concerns how to mimic the process that a human assesses the quality of sounds. This process is shown in Figure 1. In this figure, the auditory processing part contains the processing both in the auditory periphery and in auditory cortex.

However, there is one problem in modeling the human auditory processing. The knowledge about the signal processing in the auditory cortex is much less than in the auditory periphery, though there are several discoveries in neurology. Recently, psychological experiments [24] [17] suggest that joint spectro-temporal modulation of the spectrograms of sounds are important. It is also shown that humans are sensitive to joint spectro-temporal modulation of the spectrograms in low modulation rates. Several approaches that utilize the spectro-temporal modulation have been proposed.

In [27], the statistics of joint spectro-temporal modulation of envelopes are explored and these statistics vary if they are obtained from different natural sound sources. In that paper, each channel in a spectrogram was considered as a vector and a correlation matrix is obtained by cross-correlating the channels with different frequency offsets. Then the spectro-temporal modulation PSD of a spectrogram is calculated as the two-dimensional DFT of the correlation matrix.

Another method that focuses on the joint spectro-temporal modulation of the spectrogram is to process the spectrogram by a well-designed two-dimensional filterbank, which is called spectro-temporal response field (STRF) [4]. STRF
filters have respective gains, band widths, orientations and center modulation frequencies along the spectral and temporal axis to simulate responses of different neural cells. Then the outcome, which is served as an internal representation, can be seen as a stack of responses generated by convolving the spectrogram with all STRF modulation filters. This method has been used in speech enhancement [21], speech classification [23] and the evaluation of speech intelligibility [4].

As mentioned above, a non-intrusive objective assessment system can be modularized according to functionalities: the first component is the auditory processing, which tries to mimic the human perception procedure and the second one is cognitive mapping (cognitive processing), which is about how to map internal representations onto quality scores. To this purpose, some mature machine learning algorithms, such as Gaussian Mixture Model (GMM), Hidden Markov Model (HMM) and artificial neural network (ANN), can be applied. The choose of machine learning algorithm depends on the form of quality scores. The form can be MOS, DMOS or the probability distributions of opinion scores. As will be discussed later, the probability distributions of opinion scores is the choice in our system.

1.3 Assessment System Structure

The purpose of our system is to estimate audio quality in a non-intrusive manner. From the history of non-intrusive audio quality assessment, we know that most efficient systems apply a perceptually motivated auditory model. Moreover, ad hoc models, such as the vocal tract model, are limited to assessing speech quality while the perceptually motivated auditory model is suitable for all kinds of audio signals. Therefore, we designed our system to be based on an auditory model, which includes a peripheral unit and a cortex processing unit. For the cortex processing unit, the new concept of spectro-temporal modulation is particularly adopted. A cognitive processing unit resides after the auditory model to obtain an overall quality score. As we tried to simulate the complete process of a listening test, we took the effect of test environment and of devices into account. Thus, a pre-processing stage is added and the whole assessment system contains three stages in total: the pre-processing stage, the auditory processing stage, and the machine learning stage, which simulates the cognitive processing. The block diagram of the proposed system is shown in Figure 1.1.

![Figure 1.1: System diagram](image-url)

The pre-processing stage processes digital audio files, which are saved as certain formats and are presented as sounds to listeners in a subjective test. This thesis considers mainly quality assessment for narrow-band telephony applications, in which the sound reproducing device mainly refers to the headset. So we concerned only narrow-band speech signals and ignored the effects of a listening room. This stage consists of a normalization part that simulates the
volume adjustment phrase in many listening tests, which ensures that the listening level is comfortable to the subjects, and a filtering part that simulates effects of the devices used to reproduce sounds in the listening test. In our system the normalization part includes a speech level estimator described in ITU-T P.56 [30]. With regard to the loudness level that input sounds should be normalized to, it is recommended to use a high and comfortable level. The non-intrusive speech quality assessment standard ITU-T P.563 advises 79 dB SPL and this value is adopted in our system. The filtering part is to simulate the effect of a headset receiver. The characteristics of this kind of receivers are described in ITU-T P.48 [29] and a corresponding filter is implemented in ITU-T STL [31].

The auditory processing stage mimics the human auditory system and generates internal representations. As the model proposed by Dau in [6] is effective for objective quality assessment, it is adopted in our system with some modifications in order to utilize spectro-temporal modulation. Dau’s model in [6] concerns the perceptual signal processing in the auditory system and ignores those physiologically detailed processes. This model has been updated several times and the latest one is the computational auditory signal-processing and perception (CASP) model [12]. CASP model contains stages that simulate the outer and middle ear, the basilar membrane, the hair cell and the auditory nerves. The last stage is a temporal modulation filterbank, which is replaced by a spectro-temporal modulation unit in our system. The output of spectro-temporal modulation unit in the auditory model is served as an internal representation. The details on this part in our system model are introduced in the next chapter.

The machine learning stage is responsible for mapping internal representations onto appropriate scores. The auditory processing can be seen as some kind of feature extraction, for it throws away information that cannot be perceived psychoacoustically. But the output of the auditory processing stage, i.e., internal representation, is still a rather large-sized two-dimensional matrix. It is impractical to take the matrix as features directly. So it is necessary to process internal representations to expose key features. Then features are mapped onto some quantities that are related to quality. Finally according to these quantities of an audio signal and the given rating scheme, this stage outputs a score, such as MOS in ACR tests. We preferred to adopt the probability distribution of opinion scores in our system because it provides more information than MOS. Details about this stage are discussed in Chapter 3.
Chapter 2

Human Auditory Perception

As mentioned in Chapter 1, our solution requires an accurate auditory perceptual model in the auditory processing stage. The accuracy of model is vital to make results of an objective assessment coherent with that of a subjective assessment. Although lots of facts have been discovered through physiological, psychological and biophysical experiments, there are still many details that have not been studied thoroughly. Human auditory perception model develops as researchers gain deeper understandings about how human auditory system works. Most models concern about the input-output relationship of the auditory system instead of the exact biophysical mechanism. In this chapter, the human auditory system is divided into components. Each component is investigated and its corresponding perception model is discussed. Figure 2.1 shows the functional components in the auditory processing stage and the corresponding models.

2.1 Auditory System Overview

Human auditory system includes two parts: ear and central nervous system. The ear is the sensor organ that transforms sound wave from air pressure into nerve impulses. It is depicted in Figure 2.2. The central nervous system perceives impulses and performs high-level processing. The human ear has three parts: outer ear, middle ear and inner ear. The outer ear gathers sound waves in the environment. The input sound is filtered as it passes through the middle ear. As the inner ear contains liquid, air pressure, which represents sounds, changes into liquid vibrations. The cochlea, which is a part of the inner ear and is surrounded by hair cells, transforms the vibrations into nerve impulses. Those impulses finally reach central nervous system, which is referred to as the auditory cortex in this thesis.
Figure 2.1: Auditory model

Figure 2.2: Ear structure [32]
2.2 Outer and Middle Ear

The outer ear is comprised of a pinna and an auditory canal. The pinna reflects sound waves and helps humans determine the location of a sound source. After being reflected, sound waves propagate along the ear canal and hit the eardrum at the end of the canal. The outer ear filters sound waves and preserves waves in the range from 1 kHz to 9 kHz. The filter frequency response has a peak around 3 kHz.

The middle ear is an air-filled hollow cavity, which begins at the eardrum and ends at oval window. This cavity includes three ossicles called the malleus, the incus and the stapes. The main function of the middle ear is to efficiently transfer sounds with a maximal gain around 1 kHz.

The outer and middle ear is modeled as a bandpass filter with a maximal gain at 800 Hz and slopes of 20 dB/decade [12].

2.3 Inner Ear

2.3.1 Cochlea

The start of the inner ear is an oval window. The inner ear contains a cochlea, which is a coiled tube filled with liquid and has a basilar membrane all along. Air pressure that travels through the middle ear and then is transformed into vibrations of the fluid within the cochlea. There are three fluid-filled tubes in the cochlea. They are the scala vestibuli, the scala tympani and the scala media. The scala vestibuli lies above the scala media, while the scala tympani lies below the scala media. The middle tube “scala media” is also called the cochlea duct and is separated by the basilar membrane from the scala tympani.

2.3.2 Basilar membrane

Basilar membrane has a resonance structure. This means that the membrane has different resonance frequencies at different positions along it. It is shown in experiments that high frequencies resonate near the base of the cochlea (the oval window side) and low frequencies resonate near the apex of the cochlea. This property is resulted from the variance in the width and stiffness of the membrane. As shown in Figure 2.3, on the oval window side of the cochlea, the membrane is narrower than it is on the other side. The tonotopic mapping of basilar membrane is essentially spectral analysis.

Another discovery is about the relative bandwidth on various positions along the basilar membrane. The ratio between the absolute bandwidth and the center frequency approximately remains a constant along the basilar membrane. This implies that the resolution decreases as the frequency increases.

2.3.3 Hair cell

The organ of Corti locates on the basilar membrane. It is a specialized auditory sensory organ that utilizes vibration of the fluids. There are hair cells with stereocilia, which is a mechanosensing organelle and opens an ion channel to let positively charged ions flow into the hair cell as a response to the sound vibration in the fluid. Further, ions that flow into the cell trigger the neural
signals. This is how hair cells transform mechanical sound waves into electronic neural signals.

### 2.3.4 Masking Effects

The structure of inner ear causes phenomena that are referred to as masking effects. There are two kinds of masking effects. The first kind of masking effects is simultaneous masking (frequency masking). It says that the audibility of one signal can be reduced by the presence of another signal if the second signal is stronger and the frequency separation between these two signals is sufficiently small. The second and more intensive signal is often named as masker while the weaker one as maskee. The simultaneous masking effect can be explained by modeling the basilar membrane as a filter bank that contains overlapped bandpass filters. The bandwidth of these bandpass filters are described by critical bands. In anatomy, the detailed structure of cochlea also supports this filter bank model. Due to the width and stiffness variance of the coiled basilar membrane, high frequencies resonate at the base of cochlea, while the low frequencies resonate at the apex.

Early psychoacoustic experiments are under the presumption that all auditory filters have rectangular frequency response. To obtain more accurate conclusion on the frequency responses of auditory filters, some new methods are applied. One method is the notched-noise method, which suggests that the frequency responses of filters have a round-top with steep cut-offs and the bandwidth can be described by Equivalent Rectangular Bandwidth (ERB).

The conception of ERB simplifies the relationship between the filter’s center frequency and its critical bandwidth, for it treats the shape of each filter as rectangular and guarantees that an ERB passes the same amount of energy as that of the realistic bandwidth shape. The ERB in Hz is defined by the following equation:

$$ERB(Hz) = 0.108f + 24.7,$$

where $f$ means the center frequency of the filter in Hz and ranges from 100
Hz to 6500 Hz. The ERB scale represents the number of ERBs from 0 to a specific frequency and is described as the integration of the reciprocal of the ERB function (Equation 2.1):

\[ ERBs = 21.4 \log_{10} (0.00437f + 1), \]  

(2.2)

where \( f \) is in Hz. One ERB corresponds to an approximately 0.9 mm width on the basilar membrane.

Another masking effect is non-simultaneous masking (temporal masking), which means a masker appears immediately before or after the maskee. These two temporal masking are called forward masking and backward masking respectively. The temporal masking can be explained as that hair cells need time to adjust the audibility detection threshold.

### 2.3.5 Inner Ear Model

Considering the structure of the coiled cochlea duct and of the basilar membrane, different frequency components resonate at different positions along the cochlea duct. The inner ear model begins with a series of bandpass filters, which split up the signal into different frequency channels. These filters are placed equally on the ERB frequency axis, covering from 100 Hz to 4000 Hz. The bandwidth of each filter is one ERB. The auditory filter on ERB scale is described by psychoacoustic experiments and a practical digital implementation proposed in [28] is based on 4-th order Gammatone filters. After the filter bank, each channel is affected independently by envelope extraction and non-linear adaptation.

The transduction of mechanical oscillations to neural stimuli of the inner hair cells extracts the envelope of mechanical oscillations. The envelope extraction can be implemented by a half-wave rectifier that is followed by a low-pass filter with a cut-off frequency at 1000 Hz. The operation preserves fine structure of the spectrogram at low frequencies and extracts envelopes at high frequencies. Another option for envelope extraction is to use the Hilbert transform. The envelope \( A(t) \) of a real signal \( x(t) \) is the magnitude of the analytic representation of \( x(t) \), i.e.,

\[ A(t) = \sqrt{x^2(t) + \hat{x}^2(t)}, \quad \hat{x}(t) = H\{x(t)\}, \]  

(2.3)

where \( H\{\cdot\} \) denotes the Hilbert transform.

The value of the envelope is processed by an absolute hearing threshold detection before entering the non-linear adaptation stage. If the value of the extracted envelope is below the threshold, it is replaced by the threshold value [11].

The non-linear adaptation is the final stage of the inner ear model. It simulates the output of inner ear nerve and implements the temporal masking effect and Weber’s law. Adaptation means that the gain of the transfer function of the auditory system varies in response to the changes of the input intensity. According to Weber’s law, the effect of loudness adaptation in the auditory nerves is approximately logarithmical compression of the loudness. A more precise statement is that stationary signals are processed by the logarithmic compression while fast-changing signals are transformed by a linear function.

This adaptation stage is modeled as a series connection of five feedback circuits. In Figure 2.4, each circuit performs low-pass filtering and division.
The low-pass filter in each circuit has a specific cut-off frequency. The output of the low-pass filter is fed back to the denominator. The onset of signals charges the capacitor of the low-pass filter. Just after the onset, the output is a limited overshoot. If the input signal is stationary, the circuits begin to transition to the steady state after the onset. At the steady state, the relation between the output $O(t)$ of the circuit with the input $I(t)$ is $O(t) = I(t)/O(t) \Rightarrow O(t) = \sqrt{I(t)}$ and it denotes that each loop performs approximately a square-root compression. For a series connection of $n$ loops with input $I(t)$, the output of the final stage $O_n(t) = \sqrt[n]{I(t)}$. When $n = 5$, the overall compression approximates a logarithmic compression. If the input signal is non-stationary, there is not a steady stage in the circuits and the input signal is processed less compressively. For example, when the input signal is rapid varying compared with the time instants of low-pass filters, the circuits will not transition to the steady state and the value of the denominator is relatively constant. So each circuit just simply carries out a linear transform. After the offset of the signals, the capacitors begin to discharge and the time instants decide the duration of discharge. The time instants range from 5ms to 500ms to mimic the temporal masking effect.

This inner ear model processes signals from the outer and middle ear model and generates spectrograms for auditory cortex processing.

### 2.4 Auditory Cortex Processing

Compared with facts about how the ear works, less knowledge about auditory cortex processing of sound exists. However, it has been demonstrated that the envelope structure in spectrogram is important and informative. In [8], Dudley pointed out that speech and other natural sounds including music are informative low frequency waves, which are modulated by high frequency carriers. In other words, informative sounds are the envelopes of high frequency carriers and this modulation is essentially amplitude modulation (AM). Evidences from auditory physiology [16][7][26] show that neural cells in the auditory cortex are more sensitive to stimuli with joint spectro-temporal modulation.

From the mathematical view point, the basic idea of analyzing the modulation in a spectrogram is to decompose the spectrogram $S(t, f)$ into the sum of modulation components with different amplitudes $A$, spectral modu-
lation frequencies $\Omega$, temporal modulation frequencies $\omega$, and phases $\phi$, i.e., $S(t, f) = A_0 + \sum A_i \cos(2\pi\omega_i t + 2\pi\Omega_i t + \phi)$ [27], where $A_0$ is the DC component of a spectrogram and $A_i$ bears information. Thus, a spectrogram can be expressed in the form of summation of its modulation components. There are some successful methods that exploit the modulation characteristics of envelopes: MFCC focuses on the spectral modulation, the Dau’s model [5] concerns temporal modulation, Max-Gabor analysis [9] applies the two-dimensional DFT on a spectrogram and the STRF model consists of a spectro-temporal modulation filterbank.

To analyze the characteristics of modulation components of a spectrogram, there are two classical ways: the Fourier transform can describe the short-term properties of modulation components and the power spectral density (PSD) can describe the long-term properties. The PSD of a stationary signal can be obtained by applying a Fourier transform on the autocorrelation function or by averaging the squared amplitudes of DFT of signal frames. In our system, we tried two auditory cortex processing approaches. One approach is like the short time Fourier transform (STFT) and the other one is similar to the PSD analysis. Both of them process the spectrogram and utilize spectro-temporal modulation. The output of auditory cortex processing stage is processed by the feature extraction unit, which is described in Section 3.1.

2.4.1 Per-utterance Cortex Processing

This method proposed in [27] uses an autocorrelation matrix to obtain the modulation PSD. Provided that $S_i(t)$ denotes the signals of the $i$-th frequency channel in a spectrogram and $C_{i,j}(n)$ is the cross-correlation of $S_i(t)$ and $S_j(t+n)$, i.e., $C_{i,j}(n) = E[S_i(t), S_j(t+n)]$. An autocorrelation matrix is defined as the average of all cross-correlation functions $C_{ij}(n)$ with the same frequency offset $|i-j|$. This relationship can be expressed by the following formula:

$$A(n, k) = \frac{1}{M-k+1} \sum_{i=1}^{M-k+1} C_{ij}(n), \quad j = i + k - 1, \quad 1 \leq k \leq M.$$

The modulation PSD is the two-dimensional DFT of the autocorrelation matrix according to the Wiener-Khinchin theorem.

2.4.2 Per-frame Cortex Processing

In contrast to the per-utterance method, the per-frame method is more straightforward. It is very similar to STFT. The input to the auditory cortex processing unit is a spectrogram, which is produced by the preceding unit: inner ear model. First, this method multiplies a shifting window with the spectrogram and produces spectrogram frames, for instance, the $i$-th frame is labeled as $x_i(t, f)$. The length of each frame is set to 100 ms in our system. Second, the DC component of each frame is used as a feature and then it is removed from frames. The second step is to facilitate the feature extraction, for the magnitude of DC term is usually quite large and those of other modulation components are relatively small. The third step is to multiply each frame with a 2D Hanning window $w(t, f)$ and then to apply a two-dimensional DFT, i.e., $X_i(m, n) = \mathcal{F}\{x_i(t, f) \times w(t, f)\}$.

The auditory cortex processing is the end of the auditory model. The output of the auditory model is served as an internal representation, which is processed
by a machine learning stage. In next chapter, we will discuss details of the machine learning stage.
Chapter 3

Machine Learning

In this chapter we discuss the details of the cognitive processing. The previous stage is auditory processing, which transforms sounds into internal representations. This stage includes three functional units: feature extraction, utility calculation and scoring.

The feature extraction unit is responsible to pull out effective features, which are highly related to quality, from the internal representations.

The utility calculation unit simulates the procedure that human brain links features with quality. Many machine learning tools can find some nonlinear functions and perform a nonlinear mapping that transforms features into so-called “multi-dimensional utility”. The multi-dimensional utility is regarded as a set of effective quantities that are highly related to quality. In our system, we chose Artificial Neural Network (ANN) as the machine learning tool because it can implement almost all nonlinear mappings.

After the utility calculation unit generates the multi-dimensional utility of a piece of test signal, the scoring unit mimics the procedure people give an appropriate score. The scoring algorithm learns how to map the multi-dimensional utility onto scores according to a rating scheme, i.e., the ACR. In this way the scoring algorithm can be modified to simulate different rating schemes.

Although both the utility calculation unit and the scoring unit can be nonlinear, for the simplicity in implementing our system, all nonlinear functions exist in ANN and the scoring algorithm just performs a linear (in a generalized sense) mapping.

3.1 Feature Extraction

The main purpose of the feature extraction is to find features directly or strongly related to quality. Keeping the main purpose, another important principle when we implement extraction algorithms is that the extraction algorithm should cooperate with the previous and the following stage well. Since we have implemented two auditory cortex processing approaches (per-frame and per-utterance), here we introduce three feature extraction algorithms that are also divided into two categories.
3.1.1 Per-frame extraction

As mentioned in the Section 2.4.2, the per-frame cortex processing segments the spectrogram into frames and applies 2D DFT on each frame. From the 2D DFT of each frame, we extracted features called “segmental features”. Then we generated a “global feature” vector for the utterance by analyzing segmental features. In our system, we only concerned the amplitude of the 2D DFT of frames and we utilized the symmetry property. We named the amplitude of the 2D DFT of a frame of the spectrogram “modulation spectrum” in this chapter.

Algorithm 1

The first algorithm is a straightforward and plain one, which considers elements in the modulation spectrum as segmental features directly. Since the modulation spectrum is symmetric to the origin, we can throw away the third and fourth quadrants in this algorithm. To reduce the size of the modulation spectrum, the modulation spectrum of the \(i\)th speech frame denoted as \(X_i(m, n)\) is divided into smaller blocks labeled as \(B_{i,p,q}(k,l)\). Statistics of each block, such as the mean and the standard deviation, are calculated. Thus, \(X_i(m, n)\) generate smaller 2D matrixes. For example, a matrix, \(R_i(p, q)\) where an arbitrary element \((p, q)\) contains the mean of the corresponding block \(B_{i,p,q}(k,l)\). \(R_i(p, q)\) can be viewed as the features of \(X_i(m, n)\). Further, elements of the global feature could be the mean \(E(p, q)\) and the standard deviation \(V(p, q)\). They are expressed as
\[
E(p, q) = \frac{1}{N_p N_q} \sum_{i=1}^{N} R_i(p, q)
\]
\[
V(p, q) = \frac{1}{N_p N_q} \sum_{i=1}^{N} (R_i(p, q) - E(p, q))^2
\]
Finally all elements in \(E(p, q)\) and \(V(p, q)\) are rearranged as a one-dimensional global feature vector. Other segmental features can also be calculated, such as a matrix where an arbitrary element \((p, q)\) contains the standard deviation of the corresponding block \(B_{i,p,q}(k,l)\), and their statistics can be added to the global feature vector.

Algorithm 2

In our system, we concerned the telephony applications. The modulation spectrum can be seen as a combination of the modulation spectrum of three types of sounds: noise, burst and clean speech. We considered the three different types of modulation spectra as three “patterns”, as shown in the up-right subfigures in Figure 3.1, Figure 3.2 and Figure 3.3, respectively. In Figure 3.1, Figure 3.2 and Figure 3.3, the up-left subfigures depict the spectrogram and the up-right subfigures plot the modulation spectrum. As observed from the modulation spectrum, the energy distribution in a noise pattern is scattered in four quadrants but still constrained in low modulation frequencies. A burst is introduced by a strong onset of a sound, like a word or a sudden loud noise. The pattern of a burst in the modulation spectrum contains abundant strong components with different temporal modulation frequencies and very low spectral modulation frequencies. The characteristic of the burst pattern is a horizontal line located around the zero spectral modulation frequency. A clean speech pattern exhibits multiple lines with a specific orientation. The orientation is related to the change of tones.

An idea originated from the preceding observations is to relate audio quality with the shape, the intensity and the number of lines in the modulation spectrum. In image processing, the Hough transform is widely used to detect
lines. We used the Hough transform to express lines in another domain so that we could distinguish the three types of patterns of modulation spectrum more clearly. The modulation spectrum of the \(i\)-th frame, which is in the “image” domain, is transformed into a “Hough image” denoted as \(H_i(\theta, \rho)\) in the Hough domain. A line can be described as points that fall on \(y = kx + b\) in a Cartesian coordinate system. The Hough transform expresses the same line by its parameters \(b = kx - y\). In an image, a line has finite points due to the limited size of the image. With the Hough transform, an arbitrary point \((x_0, y_0)\) is mapped to a curve \(\rho = x\cos \theta + y\sin \theta\) in the Hough domain. A line \(\rho = x\cos \theta_0 + y\sin \theta_0\) with \(N\) points \((x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\) is equal to a set of \(N\) curves that intersect at \((\rho_0, \theta_0)\) in the Hough domain.

We needed to make a few modifications to the original Hough transform algorithm for detecting lines in the modulation spectrum. In fact, the original Hough transform needs a binary image as the input. The value in the Hough domain reflects the number of points with value ‘1’ lying on a specific line. To apply this algorithm, a threshold is needed to transform the modulation spectrum into a binary image, which loses the magnitude information. Our modification to the Hough transform is to maintain the magnitude of each point when transforming it into the corresponding curve in the Hough domain. Therefore, the magnitude of a point in the Hough domain reflects both the length and the strength of a line in the modulation spectrum. We defined that:

1. In the image domain, the energy of each line is the sum of energy of all points that belong to it.
2. In the Hough domain, a curve resulted from the corresponding point, which is in the image domain, has a magnitude that is equal to the energy of the corresponding point.
3. The magnitude of one point in the Hough domain should be the sum of the magnitude of all the curves that cross the point instead of the simple count of the number of curves.

A step before the Hough transform is to multiply the modulation spectrum with a circular mask, which sets the value of elements outside this mask to zero. This step ensures that the background in the modulation spectrum has the same contribution to different orientations. Otherwise, the rectangular shape of the original modulation spectrum makes the points that reflect diagonal lines to receive more influence from the background.

The three patterns of modulation spectrum result in different types of \(H_i(\theta, \rho)\). The following figures show the results of different patterns after the modified Hough transform. Multiple lines with a specific orientation (a clean speech) lead to multiple peaks at the same \(\theta\) but different \(\rho\) (bottom-left subfigure in Figure...
3.3); one horizontal line (a burst) results in a peak at \( \theta = -\frac{\pi}{2} \) (bottom-left subfigure in Figure 3.2); uniformly distributed dots (a noise) cause no dominant peaks (bottom-left subfigure in Figure 3.1).

If the size of \( H_i(\theta, \rho) \) is still too large for ANN, we needed to reduce its size. The \( \rho \) information is not so important as the \( \theta \), for we cared the orientations of lines far more than the positions. We could not sum elements along the \( \rho \) axis. If we sum elements along the \( \rho \) axis, the \( H_i(\theta, \rho) \) collapses into a one-dimensional vector. All elements of the one-dimensional vector are the same because every curve spans all \( \theta \).

We could analyze different types of \( H_i(\theta, \rho) \) further. The \( H_i(\theta_0, \rho) \) reflects the distribution of the energy of the modulation spectrum along the \( \rho \) axis at a specific \( \theta_0 \), which is between \(-\frac{\pi}{2}\) and \(\frac{\pi}{2}\). We could see that the energy distribution, \( H_i(\theta_0, \rho) \), has a few peaks in the clean speech pattern, i.e., \( \theta_0 = 0 \) in bottom-right subfigure in Figure 3.3. Other energy distributions do not have such peaks. In other words, in the clean speech pattern, there is a specific \( H_i(\theta_0, \rho) \) that is more fluctuating than other \( H_i(\theta_k, \rho) \), \( k \neq 0 \). In the noise pattern, \( H_i(\theta_k, \rho) \), \( k \in [-\frac{\pi}{2}, \frac{\pi}{2}] \) are smooth. In one word, different types of \( H_i(\theta, \rho) \) have different types of combinations of \( H_i(\theta_k, \rho) \), \( k \in [-\frac{\pi}{2}, \frac{\pi}{2}] \) and each \( H_i(\theta_k, \rho) \) has different degrees of fluctuation.

We could describe the strength of fluctuation of a specific \( H_i(\theta_0, \rho) \) by the strength of its high frequencies components. In our system, \( H_i(\theta_k, \rho) \), \( k \in [-\frac{\pi}{2}, \frac{\pi}{2}] \) are high-pass filtered and then they are element-wise squared. This operation can highlight differences among different \( H_i(\theta_k, \rho) \), for example, the values of elements of a flat \( H_i(\theta_k, \rho) \) are greatly reduced but the values of elements of a \( H_i(\theta_k, \rho) \) with dominant peaks are intensified. After this operation we could sum all elements in \( H_i(\theta, \rho) \) along the \( \rho \) axis and obtain one-dimensional vector \( T_i(\theta) \), where \( i \) denotes the frame number.

The \( T_i(\theta) \) of different patterns are shown in bottom-right subfigures in Figure 3.1, Figure 3.2 and Figure 3.3, respectively. We could see the distinguishing differences among different patterns. The \( T_i(\theta) \) of a noise pattern has no dominant peaks and is a little “chaotic”. The \( T_i(\theta) \) of a burst pattern has a dominant peak around the \( \theta = \pm \frac{\pi}{2} \) because the pattern is an approximately horizontal line. The \( T_i(\theta) \) of a speech pattern has a dominant peak around a specific \( \theta_0 \), in the figure \( \theta_0 = 0 \).

The segmental feature is chosen as one-dimensional vector \( T_i(\theta) \). Finally the global features are \( E(\theta) = \frac{1}{N} \sum_{i=1}^{N} T_i(\theta) \) and \( V(\theta) = \frac{1}{N} \sum_{i=1}^{N} (T_i(\theta) - E(\theta))^2 \). Other global features can also be added, such as the mean of the value of the DC components of all frames.

However, there are two potential risks in this algorithm.

First, there are quantization errors in the Hough domain. Due to the computer memory limit, \( \rho \) and \( \theta \) are discrete in a \( H(\theta, \rho) \). The number of rows and columns in the \( H(\theta, \rho) \) is usually not large enough to represent all possible \( \rho \) and \( \theta \). The \( \theta \) axis is usually uniformly divided from \(-\frac{\pi}{2}\) to \(\frac{\pi}{2}\). The \( \rho \) axis is uniformly divided between the maximal possible \( \rho \) and the minimal possible \( \rho \) according to the resolution configuration. The Hough transform picks valid values of \( \theta \) and then calculates the values of \( \rho \) by the Equation 3.1. The original \( \rho \) may need to be quantized to another valid \( \hat{\rho} \). The quantization gives rise to unwanted fluctuations along the \( \rho \) axis. One solution is to divide the energy, say \( E(\rho, \theta) \) to \( (|\rho|, \theta) \) and \((|\hat{\rho}|, \theta) \) proportionally, which means \( E(|\rho|, \theta) = E(\rho, \theta) \times (\rho - |\rho|) \)
Figure 3.1: Noise pattern
up-left: Spectrogram up-right: Modulation spectrum
bottom-left: Hough transform bottom-right: Features
Figure 3.2: Burst pattern
up-left: Spectrogram up-right: Modulation spectrum
bottom-left: Hough transform bottom-right: Features
Figure 3.3: Speech pattern
up-left: Spectrogram up-right: Modulation spectrum
bottom-left: Hough transform bottom-right: Features
and \( E_{\rho} = E_{\rho} \times ([\rho] - \rho) \).

Second, the resolution of \( \rho \) changes at different \( \theta \). The modulation spectrum is also discrete and has finite size. It means the \( x \) and \( y \) in the Equation 3.1 are discrete and have a limited range, for example \( 1 \leq x \leq M, \quad 1 \leq y \leq N, \quad x, y \in \mathbb{N} \). Consider such a case where all elements in the modulation spectrum have the same value. When \( \theta = 0, \ \rho = x \), so \( \rho \) has \( M \) possible values as \( x \), all points in the modulation spectrum project their energy onto \( M \) points in the vector \( H(0, \rho) \). By contrast, when \( \theta = \frac{\pi}{4}, \ \rho = x + y \) and \( \rho \) has \( M + N - 1 \) possible values according to the combination of the values of \( x \) and \( y \). Thus all points in the modulation spectrum project their energy onto \( M + N - 1 \) points in the vector \( H(\frac{\pi}{4}, \rho) \). Supposed the sum of the energy of all points in the modulation spectrum is \( E \), all \( M \) points in the \( H(0, \rho) \) have the same value \( \frac{E}{M} \), which is higher than the value \( \frac{E}{M+N-1} \) shared by \( M + N - 1 \) points in the \( H(\frac{\pi}{4}, \rho) \). In other words, the energy distribution is more concentrated in the \( H(0, \rho) \) than in the \( H(\frac{\pi}{4}, \rho) \).

These two risky problems limit the performance of Algorithm 2.

### 3.1.2 Per-utterance extraction

In contrast to the per-frame extraction algorithm, the per-utterance extraction algorithm has the input in the form of only one modulation PSD instead of a series of modulation spectra. There are several potentially effective features proposed in [27]. Some features are suitable in our system, for example, singular values, the asymmetry between the upper left and the upper right quadrant, the ratio between the power of the lower and the higher frequency components, the ratio between the DC power and the rest power. It works well in classifying different kinds of natural sounds (Figure 3.4) and in estimating SNR when the speech is affected by an additive white noise (Figure 3.5).

However, this feature extraction algorithm does not perform well in our final test due to the per-utterance cortex processing. A fact is that the average operation in the per-utterance cortex processing reduces the performance of the per-utterance feature extraction algorithm. The operation aims to reduce the high-dimensional cross-correlation matrix into an autocorrelation matrix. But differences in small envelope structure, which represent delicate quality degradation or enhancement, are also averaged by this operation.

Another fact is that the degradation noise introduced by speech codecs in lots of cases has complex characteristics in time and/or frequency domain. In some cases, statistical characteristics of the modulation PSD of noises are very similar to those of clean speeches. It is hard to evaluate the speech quality when such kinds of noise are added to clean speeches.

### 3.2 Artificial Neural Network

Aiming to map features to the multi-dimensional utility, our system adopts ANN to implement nonlinear functions. The ANN is designed to be a feed-forward network, which contains three layers of neurons. We chose logistic sigmoid function as the transfer function for both all layers. The ANN is trained to reach the mean square error (MSE) between the network output and the discrete
Figure 3.4: Modulation PSD contour of different signals
(50%, 60%, 70%, 80%, 90% power contour)
A: Babble noise, B: White noise, C: Speech, D: Music
Figure 3.5: Modulation PSD contour of a speech in white Gaussian noise of different power (50%, 60%, 70%, 80%, 90% power contour)

A: Noise (SNR = −∞ dB), B: Speech+Noise (SNR = 0 dB),
C: Speech+Noise (SNR = 7 dB), D: Speech+Noise (SNR = 14 dB),
E: Speech+Noise (SNR = 21 dB), F: Clean speech (SNR = ∞ dB)
probability distribution of opinion scores. After training, the output layer is replaced by a scoring function.

3.3 Scoring

In this section, we introduce the last stage of the proposed system and the final output of the system. According to the quality of test signals, the stage aims to give quality scores on a specific rating scale, which depends on the listening test that the system tries to mimic or other system requirements. A non-intrusive objective quality assessment system usually produces quality in the form of MOS, which is defined as the expectation of the individual opinion scores \( S_n \)

\[
MOS = \frac{1}{N} \sum_{n=1}^{N} S_n. \tag{3.2}
\]

To provide more information, a better choice is to generate the probability distribution of opinion scores. The probability distribution of opinion scores can also be used to estimate the MOS in ACR tests if we model each individual opinion score as a discrete random variable and assume that all scores are independently identically distributed (i.i.d.). Provided that discrete opinion scores have \( M \) possible discrete values, according to the definition of expectation, the MOS can be written as

\[
MOS(x) = \sum_{i=1}^{M} i \times p_{i|x}(i|x), \tag{3.3}
\]

where \( p_{i|x}(i|x), i = 1, 2, ..., M \) is the probability distribution of opinion scores. The estimated probability distribution of opinion scores generated by our system is denoted as \( \hat{p}_{i|x}(i|x), i = 1, 2, ..., M \). Further, the probability distribution of the opinion scores can be modeled by a multinomial logistic function [1]:

\[
p_{i|x}(i|x; w) = \frac{e^{w_i^T x}}{\sum_{j=1}^{M} e^{w_j^T x}}, \quad i = 1, 2, ..., M, \tag{3.4}
\]

where \( x \) is the feature vector and \( w_i \) is a weighting vector. The best weighting vectors \( w_i \) in Equation (3.4) must be found to reach the minimum classification error, in other words, to minimize some distance between estimated probability distributions and the true probability distribution.

In this system, the distance we chose is the Kullback-Leibler (KL) divergence. KL divergence is defined as

\[
KL = \int_x f_x(x) \sum_i p_{i|x}(i|x) \log \frac{p_{i|x}(i|x)}{\hat{p}_{i|x}(i|x; w)} d(x)
\]

\[
= const - \int_x \sum_i f_{i,x}(i,x) \log \hat{p}_{i|x}(i|x; w) d(x). \tag{3.5}
\]

To minimize the KL divergence is equivalent to maximize

\[
\int_x \sum_i f_{i,x}(i,x) \log \hat{p}_{i|x}(i|x; w) d(x) = E_{i,x}[\hat{p}_{i|x}(i|x; w)]. \tag{3.6}
\]
A good approximation of Equation 3.6 is shown in [3],

\[ E_{J,Y}[E_{I,X}[^{\hat{p}_{I|X}(i|x; \hat{w}(J,Y))}]] \],

(3.7)

where \((J, Y)\) and \((I, X)\) share the same probability distribution and \(\hat{w}(J, Y)\) is the maximum-likelihood estimation from the observations of \((J, Y)\). It can be proved that (3.6) can be further approximated by

\[ \sum_{n=1}^{N} \log \hat{p}_{I|X}(i_n|x_n; \hat{w}_{ML(L)}) - L, \]

(3.8)

where \(\hat{w}_{ML(L)}\) denotes the maximum likelihood estimate with \(L\) non-zero elements in it. The first term is the maximum log-likelihood subject to the constraint of \(L\) non-zero parameters and it increase as more non-zero elements are allowed. The second term is a penalty for non-zero parameters for the purpose of avoiding overfitting. This KL criterion can be generalized to

\[ \sum_{n=1}^{N} \log \hat{p}_{I|X}(i_n|x_n; \hat{w}_{ML(L)}) - \lambda L. \]

(3.9)

Introducing the definition of \(L_0\) norm, i.e.,

\[ \|w\|_0 = \sum_i w_i^0 \]

(3.10)

with \(0^0 = 0\), the target function, which is to be maximized, for our multinomial logistic regressor, is defined as

\[ \sum_{n=1}^{N} \log \hat{p}_{I|X}(i_n|x_n; w) - \lambda \|w\|_0. \]

(3.11)

This logistic regressor has already been implemented by Sound and Image Processing Lab, KTH, Sweden.
Chapter 4

Evaluation

4.1 Test Setup

A detailed system block structure is depicted in Figure 4.1. Speech files are transformed into internal representations first, then the machine learning stage adopts different features extraction methods and finally outputs estimated probability distribution.

To fulfill the training of the machine learning stage and the test of our system, we needed a database that contains noisy speech files and their opinion scores from ACR listening tests.

The noisy speech database is “Supplement 23 to the ITU-T P-series Recommendations”, which contains speech files spoken by males and females in different languages. These files are sampled at 16 kHz and used in three types of listening experiments. Speech files are processed under various degradation conditions and published respectively by AT&T (USA), CNET (France), CSELT (Italy), Nortel (formerly BNR, Canada), and NTT (Japan). This database includes two experiments that provide scores in an ACR listening test.

According to the contributor and the experiment, ITU-T P. Supplement 23 can be divided into seven subsets that contain ACR test results. This seven subsets provide 1328 utterances in total. We chose each of the subsets for testing purposes when the remaining subsets are used for training. For the $k$-th file processed under degradation condition $m$ in the $i$-th subset $D_i$, its probability distribution $p_{i,m,k}(n)$, $n = 1, 2, \ldots, 5$, where $n$ indicates the score level in an ACR test, is the normalized histogram of the scores the file receives and its MOS is calculated by definition as $\text{MOS}_{i,m}(k) = \frac{1}{5} \sum_{n=1}^{5} n \times p_{i,m,k}(n)$. Our assessment system outputs a vector $\hat{p}_{i,m,k}(n)$, which predicts corresponding $p_{i,m,k}(n)$. The estimated MOS is $\overline{\text{MOS}}_{i,m}(k) = \frac{1}{5} \sum_{n=1}^{5} n \times \hat{p}_{i,m,k}(n)$. A conditional MOS, which averages all MOS under the same degradation condition $m$, is defined as $\overline{\text{MOS}}_{\text{cond},i,m} = \frac{1}{N_m} \sum_k \text{MOS}_{i,m}(k)$, where $N_m$ denotes the number of files under the condition $m$ in the subset $D_i$ and $N_m$ is equal to 4. Similarly, $\overline{\text{MOS}}_{\text{cond},i,m} = \frac{1}{N_m} \sum_k \overline{\text{MOS}}_{i,m}(k)$. Following the convention, we chose the Pearson correlation coefficients between MOS and $\overline{\text{MOS}}$ that is fitted to MOS with a 3-rd order monotonic polynomial, to evaluate the performance of our system. The motivation for the use of the polynomial fitting is to reduce the influence of preferences of individual subjects and of the context of
Figure 4.1: System structure
4.2 Results

In this thesis we tested three methods. Method 1 uses the Algorithm 1 in Section 3.1.1. Algorithm 1 divides the modulation spectrum into small blocks, extracts segmental features from blocks and then the statistics of the segmental features are grouped as a global feature vector. In this method, the segmental features of the \( i \)-th frame of a speech signal are chosen as the sum of the each block and the value of the DC component of the \( i \)-th frame. The segmental features are rearrange as a vector denoted as \( F_i(x) \). The global feature vector consists of the mean and standard deviation of segmental feature vectors.

Method 2 also uses the Algorithm 1 in Section 3.1.1. Method 2 takes the value of the DC component of the \( i \)-th frame into the segmental features. For other segmental features, unlike Method 1, Method 2 chooses the mean and the standard deviation of each block instead of the sum. The global features are the mean and standard deviation of segmental feature vectors as mentioned in Method 1.

Method 3 utilizes the Algorithm 2 in Section 3.1.1. It uses features based on high-pass filtering of the Hough transform on the modulation spectrum. Here the normalized cut-off frequency of the high-pass filter is set to \( \frac{\pi}{2} \). On the \( i \)-th frame, the Hough image, \( H_i(\theta, p) \), is high-pass filtered along the \( p \) axis, element-wise squared and summed along the \( p \) axis. Thus, the Hough image collapses into one-dimensional vector \( T_i(\theta) \) that is taken as the segmental features. The global features are the mean and standard deviation of \( N \) segmental feature vectors as mentioned in Method 1.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Method 1</th>
<th>Method 2</th>
<th>Method 3</th>
<th>P.563</th>
<th>SIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNR-X1</td>
<td>0.7663</td>
<td>0.7949</td>
<td>0.7930</td>
<td>0.911</td>
<td>0.8692</td>
</tr>
<tr>
<td>CNET-X1</td>
<td>0.9008</td>
<td>0.8746</td>
<td>0.8800</td>
<td>0.798</td>
<td>0.8042</td>
</tr>
<tr>
<td>NTT-X1</td>
<td>0.8259</td>
<td>0.8151</td>
<td>0.8384</td>
<td>0.867</td>
<td>0.8645</td>
</tr>
<tr>
<td>BNR-X3</td>
<td>0.8688</td>
<td>0.8450</td>
<td>0.8419</td>
<td>0.923</td>
<td>0.8150</td>
</tr>
<tr>
<td>CNET-X3</td>
<td>0.7895</td>
<td>0.8232</td>
<td>0.7282</td>
<td>0.888</td>
<td>0.8978</td>
</tr>
<tr>
<td>CSELT-X3</td>
<td>0.9013</td>
<td>0.8983</td>
<td>0.8097</td>
<td>0.902</td>
<td>0.8937</td>
</tr>
<tr>
<td>NTT-X3</td>
<td>0.8677</td>
<td>0.8481</td>
<td>0.8829</td>
<td>0.843</td>
<td>0.9239</td>
</tr>
<tr>
<td>Mean</td>
<td>0.8458</td>
<td>0.8427</td>
<td>0.8191</td>
<td>0.876</td>
<td>0.8669</td>
</tr>
<tr>
<td>Variance</td>
<td>0.0028</td>
<td>0.0013</td>
<td>0.0006</td>
<td>0.0019</td>
<td>0.0019</td>
</tr>
</tbody>
</table>

As mentioned above, we investigated the performance of our system on each of the database subsets while the other subsets were used for training. Table 4.1 shows the Pearson correlation coefficients of the fitted per-condition MOS using different methods when various subsets are used for testing. It also contains the mean and the variance of the results of different methods. We can see that Method 1 and 2 outperform Method 3. This may be due to the two practical
problems of the Hough transform mentioned in Section 3.1.1. However, the Hough transform is an interesting method that deserves more investigation.

P.563 is a successful standardized non-intrusive objective speech quality assessment method proposed in [19]. It achieves a better average result than our system under ITU-T P.Sup23 database. The overall structure of P.563 is shown in Figure 4.2.

![Figure 4.2: P.563 overall structure [19]](image)

The pre-processing block in Figure 4.2 starts with an IRS filter as our system does. In addition to the IRS filter, P.563 performs voice activity detection that does not exist in our system. After the “pre-processing” block, P.563 first analyzes a degraded signal using a vocal tract model and then obtains vocal tract parameters (VTP) and LP coefficients (LPC). In the “speech Reconstruction and Full-Reference Perceptual Model” block, a quasi clean reference signal is synthesized according to VTPs and LPCs. Then P.563 utilizes an intrusive auditory model to evaluate the difference between the quasi clean reference signal and the degraded signal. The intrusive auditory model is modified from the ITU-T P.862 (PESQ). In the “Distortion-Specific Parameters” block, P.563 focuses on detecting specific degradations that happen in narrow-band telephony channels, such as noise and temporal clipping.

Regarding why P.563 has a better performance, one explanation is that P.563 is particularly designed for narrow-band telephony applications. We can see from P.563’s structure that it detects specific degradations for special cases and extreme conditions in narrow-band telephony network. Another explanation is that P.563 utilizes not only an auditory model but also a vocal tract model. Parameters from the vocal tract analysis and the basic voice quality calculated from the intrusive auditory model are both involved in the estimation of the quality.

Besides P.563, there is a non-intrusive audio quality assessment system that is implemented by Sound and Image Processing Lab (SIP), KTH. SIP’s assessment system also consists of three stages: the pre-processing stage, the auditory processing stage and the machine learning stage. In SIP’s system, the machine learning stage is divided into voice activity detection, feature calculation, utility calculation and scoring according to functionality. The pre-processing stages, utility calculation and scoring of SIP’s system and of our system are the same.

Regarding the auditory processing stage, the SIP’s system adopts CASP model as the auditory model. The diagram of CASP model in SIP’s system is
shown in Figure 4.3.

Figure 4.3: Diagram of CASP model in SIP’s system

Comparing the auditory model part in Figure 4.1 with Figure 4.3, we can see that:

1. In SIP’s system, the time-frequency transform from sound to spectrogram is performed by a complicated dual-resonance nonlinear (DRNL) filterbank [20]; while in our system, the transform is carried out by a single Gammatone filterbank.

2. The modulation processing in auditory cortex is modeled as a well-designed temporal modulation filterbank; while in our system, we implemented modulation PSD analysis and 2D DFT analysis.

Additionally, there is voice activity detection in the machining learning stage of the SIP’s system. The voice activity detection can segment spectrogram adaptively and extract features that indicate whether a frame of spectrogram is pitched (vowel) or unpitched (consonant) or background noise. Feature calculation and voice activity detection together provide features for the following utility calculation in SIP’s system.

In summary, some parts of the auditory model of the SIP’s system are more complicated and may mimic human auditory perception more accurately. For example, there is a DRNL filterbank and the parameters of the temporal modulation filterbank are well tuned. The machine learning stage of the SIP’s system also includes more techniques, e.g., voice activity detection.

SIP’s system and our system are not particularly designed for narrow-band telephony networks. The better result of SIP system may come from its advantages mentioned above. The advantage of our system is to utilize the spectro-temporal modulation in auditory model. But there are shortcomings in practical
system implementation. For example, the spectro-temporal modulation is more complex than temporal modulation in implementation so we need trade-off between the performance and the implementation difficulty. It will take a lot of time to study and tune the parameters of the two-dimensional spectro-temporal modulation filterbank, such as center frequencies, bandwidths.

The results of our experiment have illustrated the feasibility of our methodology. By combining some existing techniques, the performance of our system may be improved. For example, our system can utilize voice activity detection and use other utility calculation methods, such as support vector machine (SVM), instead of ANN. Besides, we can improve feature extraction algorithms. For instance, we can find ways to fix or alleviate the problems of the Hough transform.

Compared with existing non-intrusive audio quality assessment algorithms, our system have methodological advantages as follows:

1. Our algorithm puts no limit in the type of input signals. Although the narrow-band speech signal is currently concerned about, other types such as wide-band speech and music are also feasible. This is due to the generality of the auditory model and the feature extraction method.

2. The auditory model is based on spectro-temporal modulation, which has recently received much evidence to be effective in simulating human's auditory cortex processing.

3. Our algorithm splits the cognitive processing into utility calculation and scoring. This brings better coherence with human’s neural activities on one hand, and makes the system flexible with different rating strategies on the other.

4. The scoring unit outputs the probability distribution of opinion scores instead of the widely used MOS. In this way, more information about quality is provided.
Chapter 5

Conclusion

A non-intrusive objective audio quality assessment system is implemented in this thesis. This system includes an auditory model that bases on spectrol-temporal modulation analysis. Spectrol-temporal modulation analysis is implemented in two different ways. Several feature calculation algorithms are tried in this thesis.

It is shown that an audio quality assessment method can apply a systematic design based on general purposed auditory models and machine learning tools. This design simulates the procedure that a human rates a signal in a listening test. An auditory model mimics human auditory perception and an machine learning scheme mimics the cognitive procedure that a human perceives audio quality and gives a score accordingly. The peripheral processing and the cortex processing in an auditory model are both investigated in this thesis. Regarding the cortex auditory processing, spectro-temporal modulation analysis can be an efficient alternative. The use of statistics of spectro-temporal modulation spectrum on a per-frame basis results in features that are highly related to audio quality. Features directly extracted from spectro-temporal modulation spectrum perform better than features extracted from the Hough transform. Although the modified Hough transform in our system has several potential problems, the application of the Hough transform in feature extraction is interesting and requires further studies.

It is observed that the performance of our system does not exceed that of some existing non-intrusive objective assessment systems under the ITU-T P.Sup23 database. One reason is that our system in this thesis is not designed particularly for narrow-band speech signals. Our system puts no limit in the type of input signals. Future work could focus on tuning the auditory model, adding new techniques, such as voice activity detection, and extracting effective features for narrow-band speech signals. Some modifications to the auditory model are needed, e.g., well-tuned parameters for modulation filterbank and DRNL filterbank. The Hough transform needs more improvements. To prevent losing the generality of our system, our system should not limit to features of narrow-band audio signals, but should also take other types of audio signals, e.g. music, into consideration. Our system may choose features according to the type of the input audio file in the future.
Bibliography


