UNIVERSITÉ DE SHERBROOKE
Faculté de génie
Département de génie électrique et de génie informatique

Human Auditory Cortex inspired Audio Codec

Mémoire de maîtrise
Spécialité : génie électrique

Matin Azadmanesh

Jury : Éric Plourde (directeur)
Roch Lefebvre
Philippe Gournay

Sherbrooke (Québec) Canada July 2016

RÉSUMÉ

The attempt of this works is to incorporate modern signal processing and machine learning on one hand and new discoveries about human auditory cortex and neural coding on the other hand, in audio codec design.
A new parametrization of a typical audio codec is presented and a set of sub-optimal parameters is learned through neural networks. The error message is calculated based on human auditory model.

Mots-clés : Human Audio Perception, Neural Networks, Autoencoder, Audio Codec

REMERCIEMENTS

Consultez le *Protocole de rédaction et de dépôt aux études supérieures* de la Faculté de génie de l'Université de Sherbrooke afin de connaitre les détails sur les remerciements.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LEXIQUE

Ceci est un exemple de lexique (glossaire).

Terme technique	**Définition**
Actionneur	Définition du terme actionneur
Capteur	Définition du terme capteur
Référentiel	Définition du terme référentiel

LISTE DES SYMBOLES

Ceci est un exemple de liste des symboles.

Symbole	**Définition**
$\dot{[\,]}$	Dérivée première selon le référentiel inertiel
$\ddot{[\,]}$	Dérivée seconde selon le référentiel inertie
a	Accélération
m	Masse
t	Variable temporelle
...	...

LISTE DES ACRONYMES

Ceci est un exemple de liste des acronymes.

Acronyme	**Définition**
OIQ	Ordre des ingénieurs du Québec
UdeS	Université de Sherbrooke
...	...

CHAPITRE 1

INTRODUCTION

Audio codecs are one of the main modules of many of digital systems dealing with audio. These algorithms have been designed to transform the audio signal to a bit stream and do the reverse transformation, from bits to audio. The main objective of the algorithm is fidelity to the original signal. Keeping the bit rate (length of the bit stream) and computational expensive of the transforms as small as possible are also desirable.

Since codec design is an old problem, there has been proposed a large set of techniques to address this problem [**?**], [**?**]. Despite the importance and complexity of the problem, there has been a narrow set of techniques which has been used repeatedly in different systems. Whereas the objective of an audio codec is to keep the similarity of input and output *for human perception* ; in most of commercial audio codecs, there are several blocks and solutions to enhance the perceptual quality of the output. Main tools for quality enhancement consists of bit budgeting and perceptual masking. These methods are based on empirical results in psychoacoustics.

The objective of this work is to address these issues. The first objective of this work is to, incorporate new methods and tools from machine learning. Machine learning was the center of attention in signal processing community and beats the results of other methods for recognition related tasks. Deep neural networks were the super star in the different application stages. One of the interpretations of the deep learning is *representation learning*, i.e. instead of using an off-the-shelf feature extractors, the model learns the (sub-) optimal algorithm to extract proper feature for the specific task.

Autoencoders share some structural similarities with our design problems. A known neural networks training algorithm such as a version of backpropagation can optimize the modified autoencoder parameters and represent the codec.

The second objective of the current work, is to develop and utilize a model for human perception of audio quality. Different model for human auditory system have been proposed. It is presumed that by the aim of new findings in neuroscience and functions of auditory cortex, better approximation of human perception of audio may lead to design better codec. There are different possible paths in design of functional human brain inspired assessment system. There are simplified models of each step of auditory system from the external ear to the high auditory cortex. A smooth, i.e. continuous and differentiable model (function, stochastic process, etc.) can serve as the basis to calculate the error signal

and update the parameters regards to that error signal.
Next chapter of the current document provides a review over techniques and methods that so far have been used in commercial codecs. Next a model of human auditory system will be addressed. In the last chapter the autoencoder based codec and the primary results of the model will be reported.

CHAPITRE 2

CCEPTION

CHAPITRE 3

RÉALISATION

3.1 Introduction

We design a complete codec system using a feed forward neural network. The model is being trained the autoencoder while each number in the code has been rounded up to the closest centroid.

3.2 Introduction and Background

Codecs are algorithms consist of two parts, encoder and decoder. The Encoder transforms the information signal (audio, video, image, etc.) to a bit stream. The decoder map the bit stream to the original signal space. The main challenge in designing a codec is to keep the quality while remaining the bit rate low.
In this work the codec is being designed by learning the weight of a feed forward neural network. We simplify the the codec by the autoencoder associated with a vector quantizer. Autoencoders are multi-layer feed forward neural networks which are used to initialize to the neural nets for other (like classification, etc.) tasks. Generally autoencoders use weight sharing techniques to reduce the complexity and number of parameters, i.e. the networks is divided to the symmetric sub networks. Weight matrix are transpose of each other in corresponding layers. We call the second half of the network *decoder* and the first half followed by a vector quantizer system the *encoder*.
Vector quantizer is a algorithm seeking for centroids as density points of nearby lying samples and represent each point by its closest centroid. The set of all centroids is known as *code-book*. The code-book can be constructed by k-means algorithm and its cardinality is a pre-determined parameter in this model. See Figure 3.1). Similar ideas have been visited in [**?**] in general and in [2] for spectrogram of the audio signal.

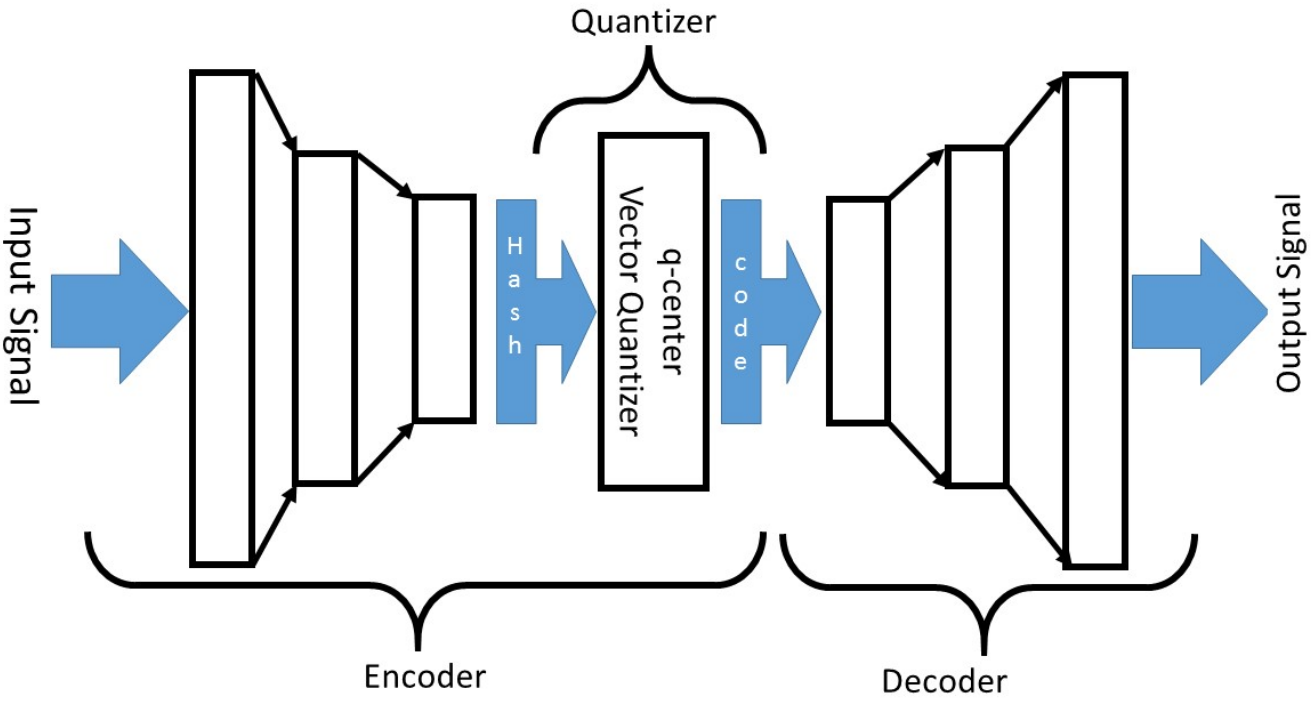

Figure 3.1 Block diagram of the Codec based on stacked autoencoders

3.3 Experiments

We run the experiment on the "mnist" dataset. 60000 samples in form of vectors, size 784×1. The data has been normalized between 0 and 1 beforehand.
We conduct different experiments by changing the number of layers and number of units in each layer. We kept the number of neurons just before and after vector quantization layer the constant number of 256 units. The transfer function fixed to *sigmoid*, $y = \frac{1}{1+e^{-x}}$ for all the layers. the hidden layers in different experiments posses variant values of the set $\{128, 256, 512, 1024\}$ and number of hidden layers varies from 2 to 4 in different experiments.
At the beginning of every epoch, Vector quantization algorithm finds q centroids among the outputs of the encoder (hash vectors) and replace each hash vector element by its corresponding (closest) centroid in the code-book.
The objective function is the Euclidean distance between the input and the output signal (waveform matching approach). The training has been done using backpropagation algorithm. We have used Optimizer that implements the Adam algorithm. It is a gradient-

based optimization based on adaptive estimation of moments. This algorithm does not perform worst than the other competitors.

3.4 Results

The main parameter we sweep over and monitor the results is the effect of the number of centroids over the Signal to Noise Ratio (snr) which calculated based on l^2 distance between the original signal and the reconstructed signal. As we expected, generally snr will increase with more centroid, the snr improvement rate is proportional to inverse of the number of centroids. See Figures 3.2 and 3.4.

The main distortion caused by the coding system and the quantization mechanism is visible as blurred edges and some low power confusions on the smooth areas of the image.See Figures 3.2 and 3.4.
The behavior of snr curve remains quit similar in different trials. We conducted various experiments on a variety of network parameters. These parameters consist of basic structure defining parameters like number of hidden layers and number of units per each layer and training parameters like batch size, learning rate, and maximum epoch.

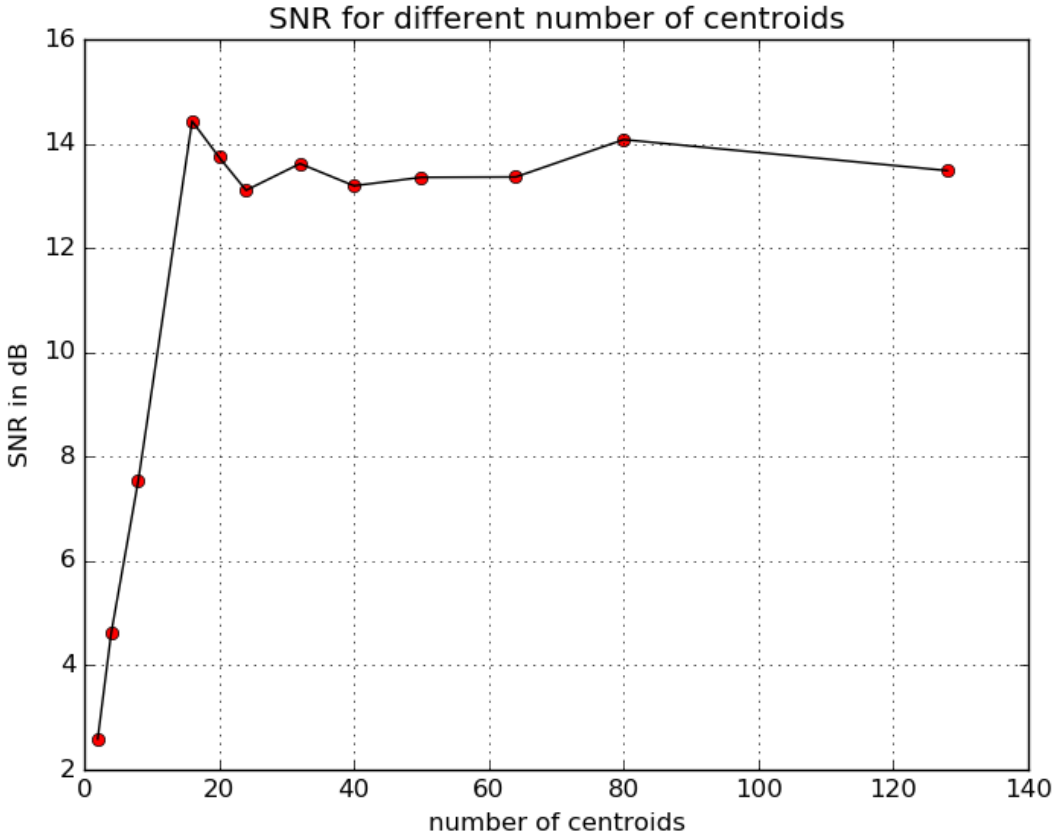

Figure 3.2 SNR different number of centroids for network {1024, 512, 256 }

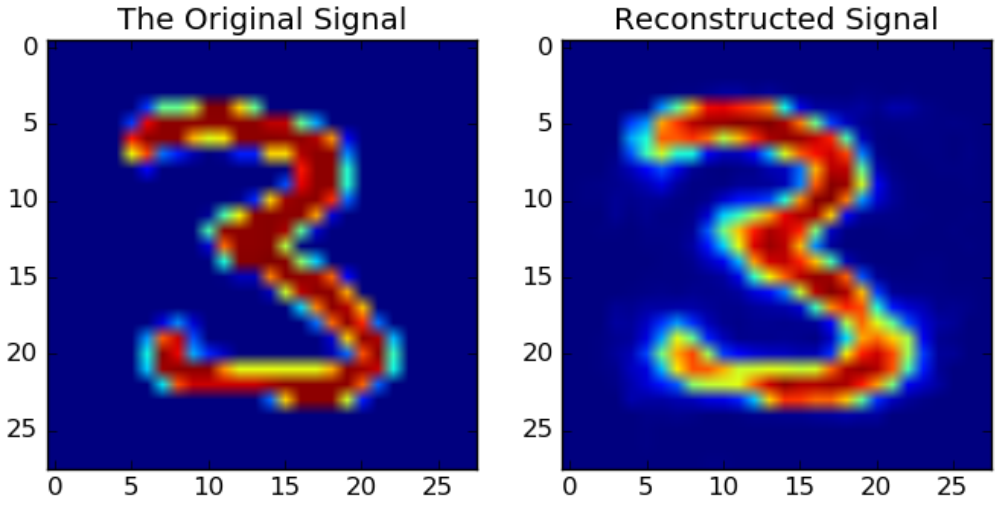

Figure 3.3 A sample Input and output of the network{512, 1024,256 } and 20 centroids

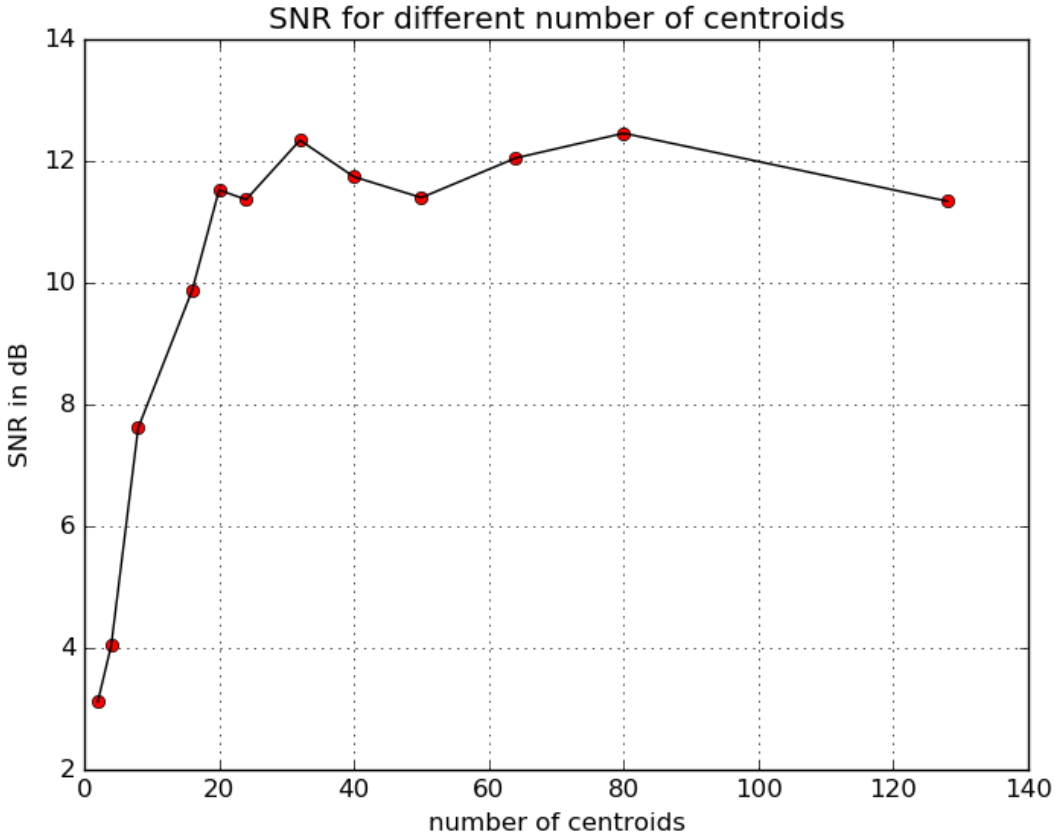

Figure 3.4 SNR different number of centroids for network {512, 1024,256 }

3.5 Discussion

The similarity of snr curves and its saturation behavior for large number of centroids is can be result of several causes. We hypothesize that this saturation may reflect the learning capacity of this network with that particular learning algorithm. The other hypothesis states that this estimator is biased and there will be some level of distortion on reconstruction for this input signal given the quantized observations.

3.6 Future works

We will examine different network structure (deeper networks), different units (convolutional, LSTM, etc.), different training algorithm and its effect on the performance of the model, and coding ability of this model for different data modulities such as audio in transform domain and time domain. Prove or reject of the hypothesis provided in section before is also the question of interest from theoretical view point.

LISTE DES RÉFÉRENCES

[1] Atreya, Anand and O ?Shea, Daniel. *Novel Lossy Compression Algorithms with Stacked Autoencoders.*

[2] Deng, Li and Seltzer, Michael L and Yu, Dong and Acero, Alex and Mohamed, Abdel-Rahman and Hinton, Geoffrey E. *Binary coding of speech spectrograms using a deep auto-encoder.* Interspeech,1692-1695,2010,Citeseer.

CHAPITRE 4

TESTS

CHAPITRE 5

ANALYSE

CHAPITRE 6

CONCLUSION

6.1 Introduction

There are different audio coding techniques that each of them performs its best for a limited set of input signal (e.g. speech or audio) and a certain bit rate range. Providing a comprehensive survey of audio codecs is not in the horizon of this text.

6.2 Classification of Audio Codecs

There are different classifications of audio codec. Forming a new paradigm for audio coding is the promise of this document, the classification by coding techniques is provided here. This classification is based on (

6.2.1 Parametric Coders

This codec is designed to provide intelligibility in very low bit rates. Parametric codec assumes a model for production of audio signal. Encoder extract the parameters of model and decoder produce the signal by substituting the received parameters in the same model. According to the assumptions about signal production, this codec performs well only for a specific class of signals, e.g. speech signals.

6.2.2 Waveform Coders

These algorithms try to minimize the distance between input and output signals, so higher bit rate is required.

6.3 Hybrid coders

These codecs advantage from both parametric codecs by taking a speech production model into account, and enhance the quality by using richer models for excitations. Codebook

excitation is one of these methods which has been used in CELP codec. The majority of modern codecs belong to hybrid compression systems.

6.4 Main Techniques in Audio Codecs

6.4.1 Linear Prediction

Linear Prediction Coding (LPC), was first proposed in [] in 1971 for speech coding. LPC is linear mathematical model of human vocal tract. [] Generally it is formulated by an all-pole filter. LPC filter extract spectral envelope of the speech. It assumes the input signal (audio or speech in this case) are generated according to an autoregressive model. Based on framing effect (setting samples values outside of the frame used for correlation calculation to zero) this method add some inaccuracy to estimation even for autoregressive model of the same order. LPC finds the mean square error answer to the following equation.

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1r} \\ p_{21} & p_{22} & \cdots & p_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ p_{r1} & p_{r2} & \cdots & p_{rr} \end{bmatrix} \tag{6.1}$$

$p+1$ in formulas above indicates the order of the filter. This problem can be solved for a using Levinson-Durbin algorithm in quadratic time.

6.5 Perceptual Masking

It has been known for a long time that mean square error (MSE) does not resemble perceptual similarities. The nonequivalence roots on a phenomena called "masking". Masking, models situations in which one sound overshadowed the other in human ear. It happens based on limited temporal and spectral resolution of auditory system. Masking happens when two similar signals (in spectral and/or temporal domain) present and the weaker one turns inaudible. Based on similarity of signals in temporal or spectral domain, the masking may be called temporal masking or spectral masking respectively. There is a large number of studies in modeling and simulating and implementation of masking and almost all codecs benefits from these results by dedicating less bits to inaudible component of signal. This is known as perceptual bit-budgeting. [] In [], perceptual distance between two signals has been defined as Euclidean distance between the linear mappings of the signals. This process also can be seen as applying the kernel method where the kernel is 3-dimensional time-frequency curve described in [], an estimation of this filter is shown in

2.1. This scheme of perceptual masking is very useful in perceptual coding and transform coding.

6.6 Human Auditory Model

More complex mathematical encoding-decoding scheme let us have higher quality in lower bit rate. The other important aspect of a "good" codec depends on how does it model the human auditory system and incorporate this modeling in design of a better codec. Studying and modeling of auditory system is quit old topic and almost all systems benefit from the results of these studies to some extends 2.4.5. The physiologic and biological analysis of human auditory system is outside of the scope of this work. For the sake of pragmatic reasons, only contemporary mathematical models which are applicable in the codec design process, modeling different stages of the auditory will be presented. These steps have been studied and their mathematical model were included in various existing audio systems. The early stages of the model consists of cochlear filtering, hair cell transduction, the phenomenon of lateral inhibition in the auditory neural pathway and midbrain integration. Mathematical models for auditory cortex have been relatively developed recently []. Cochlea is modeled as a filter bank of constant-Q bandpass filters hcoch(t ; f). As figure 2.3 indicates, the output of the cochlea filter bank will passed trough a model of auditory nerve pattern. Inner hair cell itself is modeled by a three-step process, first, high-pass filter tycoch(t ; f) then the nonlinearity ghc(:) (compression) and at the end a low-pass leakage filter hc(t). Cochlear-nucleus neurons are modeled as derivative with respect to frequency fyAN(t ; f) followed by a rectifier max(tycoch(t ; f) ; 0). The last step of the first section of the model is a short-term integrator midbrain(t ; f) with time constant , that models the further loss of phase-locking observed in the midbrain. [] The second part of the model which is also one of the focus point of this project is the model of auditory cortex have been developed and improved vastly in recent years by conducting experiments on humans and more on animals []. Basically the current model of assumes that audio (output of the early auditory stage) based on specific properties makes special regions of the auditory cortex fire the most. These spatial varying firing rate leads to Spectro Temporal Response Filters or STRFs. STRFs have been modeled as a redundant two to three or four dimensional mappings (based on the notation and not much difference on implication). The first models are three dimensional wavelet like filters which analyses the spectrogram of the input (audio) signal to three components called rate (temporal modulation), scale (frequency modulation) and direction (the signal is upward

in tone or downward). Further studies try to make this model more accurate by adding extra boundaries our condition to it [].

ANNEXE A
DONNÉES