Audio compression

Abstract

We discuss different aspects of audio compression, starting with several methods of compressing audio. Methods of quantization of a signal and how different psychoacoustic models can be used are described. Some areas where applicability of speech compression is high are studied, a few methods are considered. Experiments concerning speech compression are implemented, we discuss how they are performed and the results encountered.

Project Report, Signals and Systems
Fall 2002

Henrik Hofling (heho5769@student.uu.se)
Teodor Berglund (tebe7689@student.uu.se)
Axel Vaara (axva3373@student.uu.se)

Table of contents

Intro

A computer works in the digital world, a world where everything is discrete. In this world the computer processes data, bit streams of information. The information can be about anything, a command to switch on/off a light, a text file or an audio sequence.

The technique of transforming an analog signal into the digital world has been know for quite some time now, but the methods of doing this gets better every day. We now no so many things about the characteristics of the sound and how the human ear and brain works that this transformation, from analog to digital and back has become almost an art, an art with a lot of mathematics involved.

Since the invention of the CD, the Compact Disc, audio has been digitalized and played more and more frequently. The quality and the simplicity that the CD brings have greatly improved the enjoyment of sound, but today this is not enough. By the introduction of Internet and the hundreds of sound and music applications the new technology requires, new areas, and therefore new restrictions and demands have been put on the signal processing.

By making the analog sound to a digital bit stream, a whole world of opportunities opens. In the digital world, the possibilities of saving and manipulating the information are vast and productive. This is where the challenge lies, to make the signal processing as good as possible.

Audio compression

You start by digitalize the analog signal by sample it, this results in discrete values that describes the continuous analog signal. By doing this a certain amount of noise is added to the signal. This is because of the analog signal having continuous amplitude values and this must also be made discrete in the sampling process, all samples gets quantized.

Quantization

The CD has 16bits of quantization levels which makes the CD quality very accurate. By reducing this resolution the amount of data is reduced. This is not done without any side effects. By reducing the resolution the noise of the signal is increased because the samples has to be quantized more, the quantization noise is larger. To make this noise as small as possible are there several methods of making the quantization as accurate as possible with as few quantization levels as possible. The standard model is the linear quantization model where the levels are equally spread on the amplitude spectrum. If we know the characteristics of the sound another method of assigning quantization levels might be better. Tanh is a method that produces more details to the higher amplitudes with the result of less detail in the lower amplitudes. The corresponding method is the sinh method, which does the opposite.

Splitting up the signal

Splitting up the signal is possible by letting the signal pass through different passband filters. The resulting signals contains then of the different frequency spectrums of the sound. Since these different spectrums behave differently, it is now possible to modify one spectrum at a time. By using different quantization methods on the spectrums that is optimal for each, a more accurate quantization is made. After this, different psychoacoustic methods can be applied on the spectrums to reduce the signal further.

Psychoacoustic models/methods

There are several compression methods that can be used without decreasing the quality of the signal thanks to how the human brain and ear works. By using the knowledge of how the brain and ear functions, we can design different psychoacoustic methods that use the fact that some information in a signal is unnecessary for our interpretation of sounds, thus they can be removed. The analog signal contains of lots of frequencies, many of whom the human ear can't hear. By removing these frequencies from the signal, the information load gets smaller without effecting our impression of the signal.

Threshold of quiet

The human ear is hears differently well on different frequencies. In some frequencies the signal must be quite strong for the ear to notice it. If a frequency is below the threshold of quiet of the ear it can be removed. This is most significant in the low and high frequencies. The ear is much more sensible in the middle frequencies, the spectrum where the human speech lies. This has probably a historical meaning because humans needed to here these sounds better than others.

Frequency masking

Frequency masking occurs when a frequency we can normally hear is masked by a nearby frequency. The ear can not simply distinguish frequencies to close to each other. The masked frequencies can be removed.

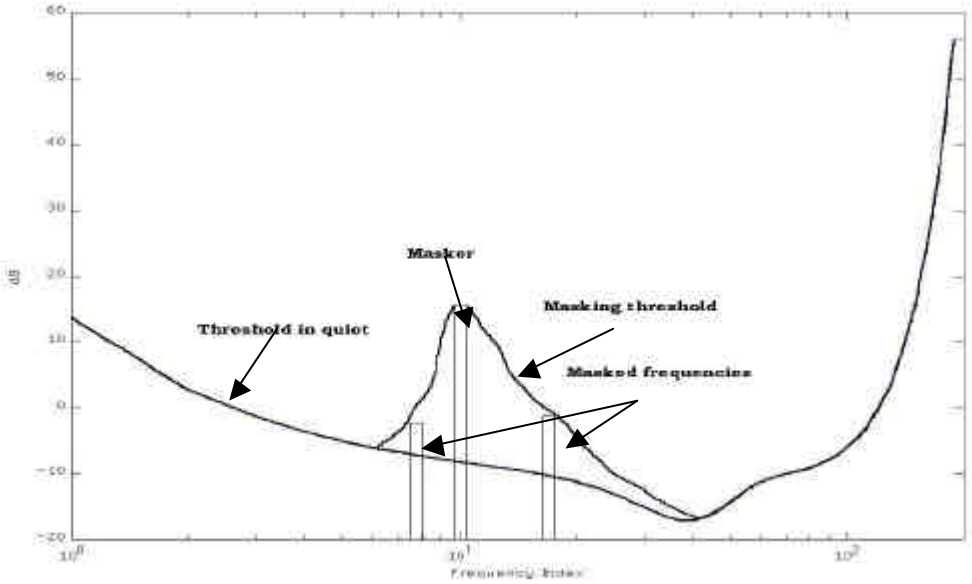


Fig.1 Graph over threshold in quiet and frequency masking

Temporal masking

When a week frequency is preceded by a strong frequency in the time domain, that is a frequency with low energy close to a frequency with high energy, the sound associated with week frequency can't be heard if the time interval between the frequencies is short. This is called temporal masking. By removing all frequencies that are masked, the ones with low energy, the information amount is reduced without effecting our interpretation of the sound.

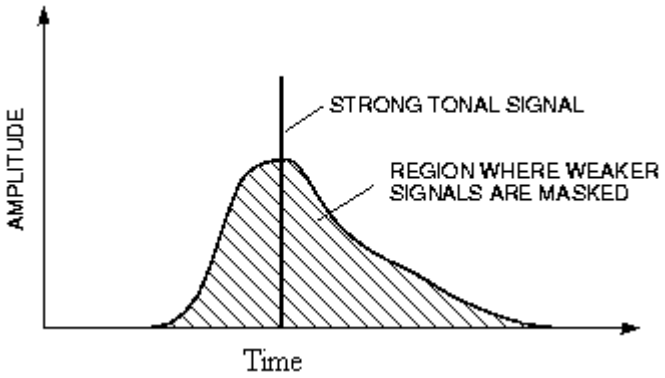


Fig. 2 Temporal masking

Areas of concern for audio compression

Although high bit rate channels and networks become more and more accessible, low bit rate coding of audio signals is more important than ever. The main reasons for a low bit rate coding are the need to minimize the transmission cost or to enable cost efficient storage, the demand to transmit data over channels of limited capacity

such as mobile networks, and to be able to store more data on less memory space such as though usage of mpeg.

Speech coding

An area where audio compression is used frequently is when there is need for speech compression. This is important when using media that require low bit rates. The most common areas of use are telecommunication networks.

In these cases there are some main concepts that are of concern when choosing an algorithm for the speech coding. The bit speed should be low, quality high, the delay small and the complexity of the algorithm should not be too big. It is not possible to accomplish a solution that is the best possible in all these four areas; we need to find a good weigh between them, depending on usage.

The LPC speech coding basis

Many speech coders have been developed over the years. The most known are based on a Linear Prediction Coding (LPC) model, and later, with extensions of or additions to –this model. The LPC has been developed with knowledge about the human vocal tracts in mind, and how they operate. With a mathematical generalized model of how the vocal tracts work it is possible to predict the next sample of a speech signal with previous samples. This is due to the fact that human speech is not independent in time. The sounds generated cannot vary completely independently, because of human physiology.

The LPC, with extensions, is implemented in modern telecommunications networks such as GSM. Also, new techniques have been researched and discovered to minimize the number of bits needed to be sent even further. The LPC however, still forms a first step for modern coders.

There are several methods already available for LPC prediction; in general, they are concerned with auto-correlation and covariance in stochastic processes, e.g. signals. There are also versions for LPC prediction based on lattice mathematics. On the receiver side, the inverse LPC-filter must have been computed for decoding.

The Analysis by Synthesis coder (GSM)

The so-called Analysis by Synthesis coding style can be used in several applications. Here, GSM is used as an example to show its applicability. GSM however, uses an extended encoding scheme based on AbS. The exact algorithm specification, down to bit level, for GSM can be found on www.etsi.fr for free.

There exist three different GSM speech coders. The first algorithm for speech coding in the GSM network was released 1987, it has been improved through the years and new clients use the newer algorithms. Older clients can still be used in the network because the base stations can still communicate using any algorithm.

The GSM network uses the Analysis-by-Synthesis (AbS) coder. The prerequisite for this type of encoding is that at the encoder side, the encoder also features a local decoder identical to the one at the receiving side of the communication, this is to be able to test synthesized speech quality.

Generally, the AbS coders work by comparing the original speech signal with a compressed version of the same signal that has been synthesized with a filter, which also may be understood from the name. The difference between the two is

then calculated to see where there is need for improvement in the next iteration. This is done in an iterative manner until we get an acceptable result. Normally, an AbS filter uses 20ms frames of the speech for each encoding, thus the coding is done on a frame-by-frame basis.

In each iteration, after each calculation of the difference, a weighting is performed. This emphasizes frequencies where speech energy is low, so when minimizing, the error energy will be focused to frequencies where the speech has high energy. This will lead to a masking of the noise by the speech, which has proven to give output enhancements in perceptual tests.

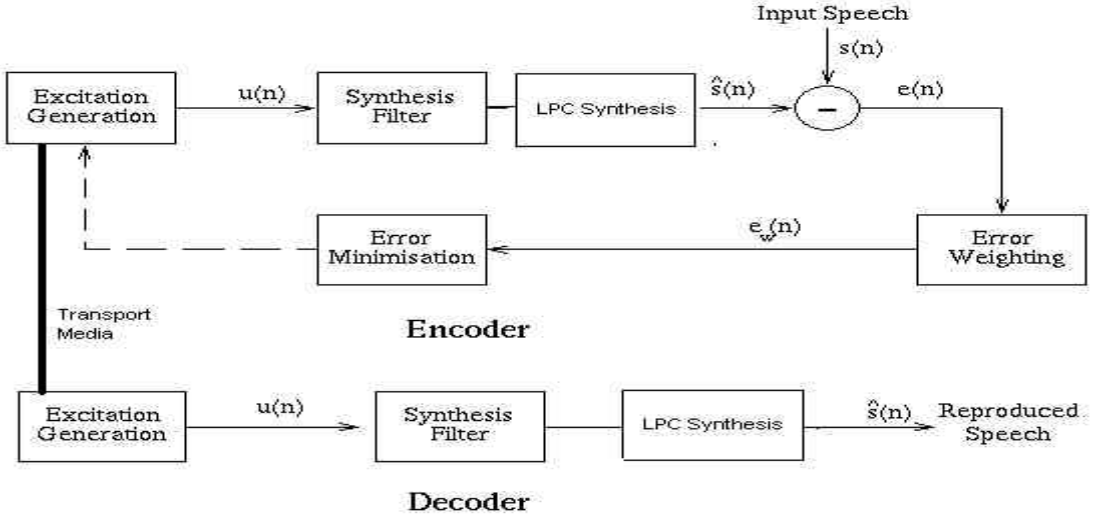
An error minimization is then performed and we derive new parameters for the filter. This means that for each iteration we minimize the error that we have computed. The optimal result is that the synthesized speech is identical to the input speech, but this is not possible in real implementations due to the time it would take. The exact error minimization method depends on the specific implementation.

After error minimization, new parameters for the filter are determined. These parameters must be transmitted to the decoder at the receiver side of the communication link.

Different methods can be used to generate the excitation signal. An adaptive codebook (CELP coders) can be implemented by storing vector samples from previous excitations. After the weighting has been performed, possible excitation signals are considered and the best is selected. With codebook matching the output can be generated faster.

There are techniques for decreasing search time in the codebook, such as using Gray-code exploration techniques. There are a number of fixed codebooks made for implementations where they are applicable.

Overview of the Analysis-by-Synthesis coding principles:



In the picture we identify the following blocks:

Input speech: The original unmodified human input speech.

- : Operator to compute the difference between the Input Speech and the synthesized speech

Error Weighting: The error vector is weighted to increase the overall performance of the error minimization procedure due to the human perceptual domain.
Error minimization: Compute new parameters for the coder that will be used to generate the excitation signal.
Excitation generator: The excitation signal that will feed into the synthesis filter on the other side of the communication is computed. Different methods can be used. In the original 1987 GSM implementation RPE (Regular Pulse Excitation) was used, we will not discuss that further here.
Synthesis filter: The coded speech is decoded and fed into the LPC for speech production.
LPC Synthesis: The Linear Prediction Decoding is performed, and output speech.

We identify the following vectors (as functions in the diagram):
s(n): Input speech signal
s(n): Reconstructed speech signal
e(n): Error vector
e_w(n): Weighted error vector
u(n): The output signal generated by the coder

Soundstorage – Mpeg(Moving Picture Expert Group)

Mpeg is becoming the standard for audio compression in an increasing amount of applications. The format consists of three different layers, three different levels of complexity. Mpeg-I/Layer 1 is the simplest thou it is still quite complex. It implements all the features described above. Mpeg-I/Layer 2 is very similar to Mpeg-I/Layer 1 in its structure, but it has a more sophisticated quantization and a more efficient scalefactor usage. Mpeg-I/Layer 3, or MP3 as it is also called, introduces many new features, in particular a switched hybrid filterbank. It also uses the analysis-by-synthesis approach. MP3 is an art in audio compression that uses many special features.

The other layers (Mpeg-II/IV) of the Mpeg standard are used for multi channel coding and for other multimedia forms such as digital video.

Implementation experiment

Background with prerequisites

There are several different implementations that can be done to compress audio data. Many methods are quite complex and includes adaptive signal processing, which wasn't covered in this course. Since the theoretical part weight a lot in this project less work has been done in this section. Speech coding is very important and fundamental in the telecommunication and telephony line of business. This project purpose partly involves these two branches and therefore the practical implementation is speech coding. The problem to solve is that uncompressed speech would take up to much bandwidth on the telecommunication and telephony media used to carry the data. It must be compressed in real time and also obtain a big compression ratio.

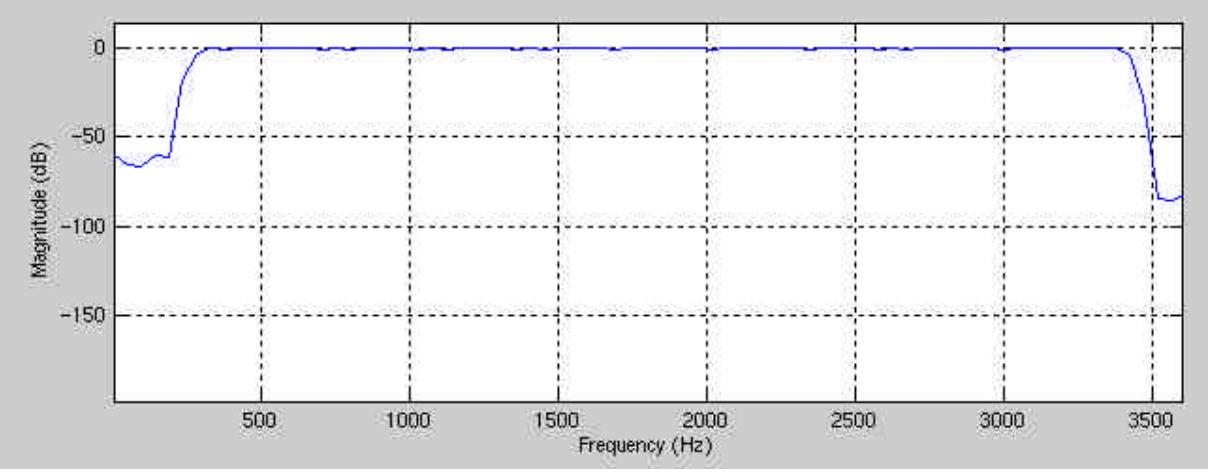
The system

Since the implementation was done in matlab the only important component is the FIR band pass filter of order 1215 with limits:

stop1 200 Hz

pass1 300 Hz
pass2 3.2 kHz
stop2 3.5 kHz

The filter:



Problem

Uncompressed audio is sampled in a rate of 44.1 kHz to make is possible to detect and recreate signals with frequencies between 20 Hz and 20 kHz. The standard quantification use 16 bits. The problem is that the uncompressed audio cannot be sent directly through the media of current interest. The audio must be compressed first.

Solution

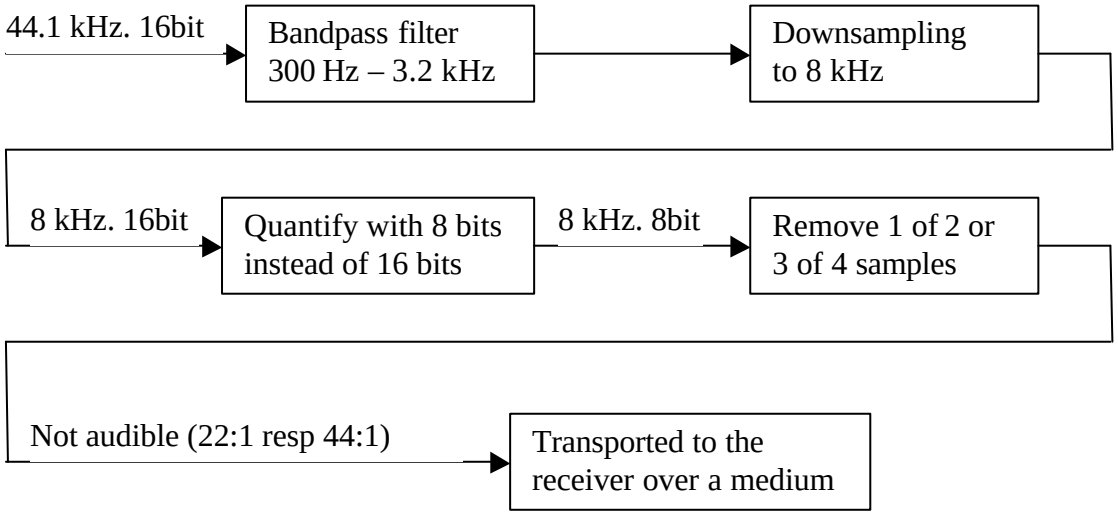
The solution is to take advantage over the fact that human speech in the frequency domain only uses frequencies between 300 Hz and 3.2 kHz. The remaining frequencies can be filtered out with a band pass filter.
The first step in the compression was to design an appropriate band pass filter and filter the signal. This was done using the filter design tool in matlab combined with the command filter.

The Nyqvist criteria demand a sampling rate of at least two times the highest frequencies in the signal. The filtered signal has an upper bound of 3.2 kHz and can be sampled at 8 kHz. The down sampling was done by selecting 1 of every (44.1/8) samples in the signal. The actual compression ratio of this operation is (44.1/8) = 5.5. The standard quantification use 16 bits, which corresponds to more then 64 000 different levels in the amplitude. This accuracy is excessive in speech coding and as a result of that 8 bits can be used for the quantification. The samples in the signal are actually just rounded off and the compression ratio is exactly 2. This combined with the down sampling compresses the data 11 times.
Further compression can be achieved by simply removing samples from the already compressed signal. Two different compression ratios were achieved by removing 1 of 2 or 3 of 4 samples. The total compression ratio after these operations is 22 respectively 44 times. The signal is not longer audible directly. The receiver of the compressed signal must interpolate the missing samples to make the signal audible. The command interp1 with the cubic interpolation method were used to do

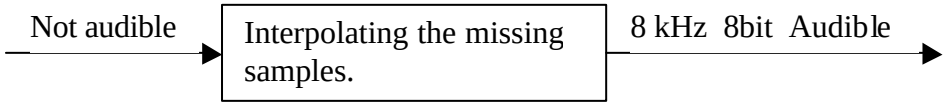
this. The 44:1 compressed signal was interpolated in two steps to achieve the best result.

Flow chart:

Audio encoder (sender)



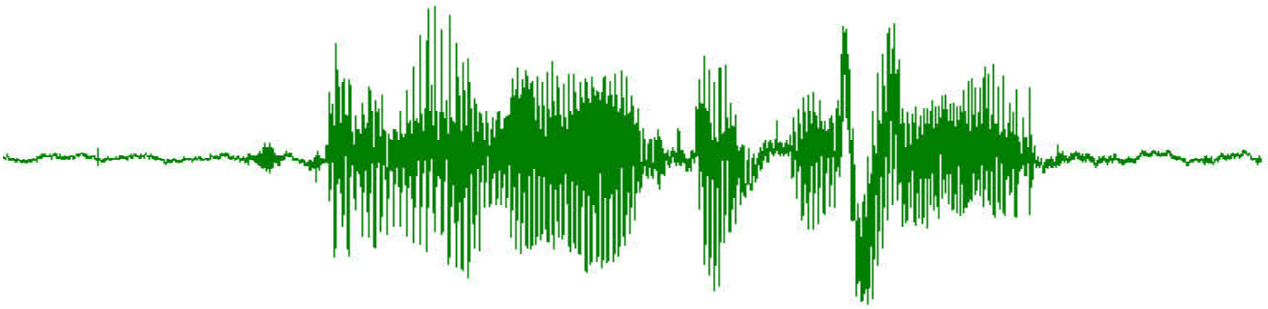
Audio decoder (receiver)



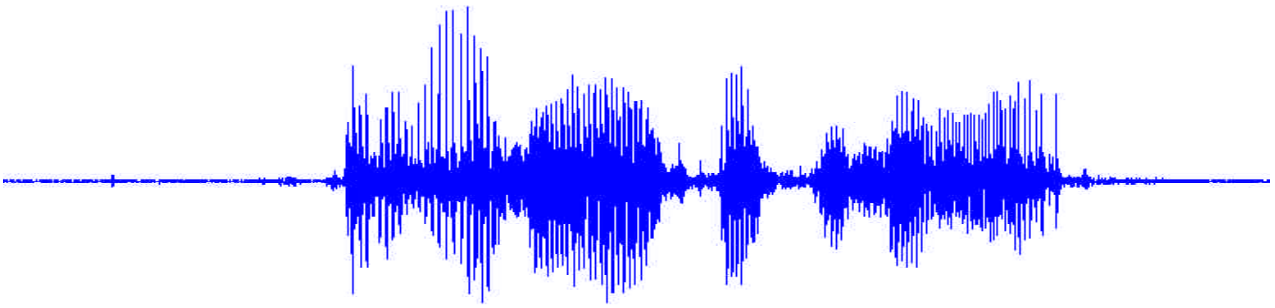
Results

The original speech is fully audible when using compression ratios up to 22 times. When compressed 44 times the speech is still audible, but with much more noise.

Uncompressed: http://user.it.uu.se/~tebe7689/uncompressed.wav

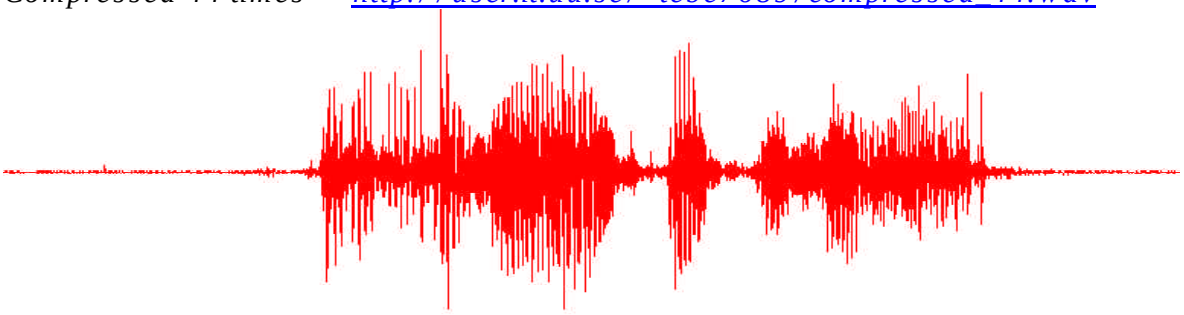


Compressed 11 times: http://user.it.uu.se/~tebe7689/compressed_11.wav



Compressed 22 times *http://user.it.uu.se/~tebe7689/compressed_22.wav*

Compressed 44 times *http://user.it.uu.se/~tebe7689/compressed_44.wav*



Conclusions

A very good compression ratio was obtained with fairly simple and fast methods. The most compressed signal is audible, but the personality of the person speaking is somehow lost and the background noise is too high for it to be useful. The quality of the second most compressed signal is acceptable and can be suitable for applications in the telecommunication and telephony branches.

Possible improvements

For storage purpose the compression can be increased more by using the characteristics in the frequency domain and apply methods like frequency masking and temporal masking. This is however costly and wouldn't be possible in a real time communication system.

References

Lecture notes by Mathias Johansson, Signals o Systems Department, Uppsala Univ. http://www.signal.uu.se/

Matlab help files

Scientist and Engineer's Guide to Digital Signal Processing:
http://www.dspguide.com/pdfbook.htm

Project homepages on Rice University of audio/speech compression:
http://www.dsp.rice.edu/courses/elec301/

Image Coding Group at Linköping University
http://www.icg.isy.liu.se/

Communications Research Group at University of Southampton
http://www-mobile.ecs.soton.ac.uk/

Faculty of Electrotechnics and Infomratics at TUB
http://iv.tu-berlin.de/

Booklets from www.engnetbase.com:

Peter Noll. “MPEG Digital Audio Coding Standards.”	2000
Davidson, G.A. “Digital Audio Coding: Dolby AC-3”	1999
Deepen Sinha, et. Al. “The Perceptual Audio Coder (PAC).”	2000
Cox Richard, “Speech coding.”, AT&T Labs	2000