Voice Compression Systems for Wireless Telephony

Javier Bustos J. - Alejandro Bassi A.
Departamento de Ciencias de la Computacion. Universidad de Chile.
{jbustos,abassi}@dcc.uchile.cl

Abstract

This work presents a comparative study between three voice compression systems for wireless telephony, CELP, VSELP and GSM 06.10, and one system based on Artificial Neural Networks (ANN). The main result is that the system based on ANN exceeds the best current performance standard (CELP), however a speaker dependency hinders its potential standardization.

1 Introduction

Voice compression is a mature technology with many commercial applications for wireless telephony. However, this area may still benefit from innovative approaches that diverge from the mainstream. To asses the relevance of a new method, a comparative study with well-established techniques is indispensable.

This work stemmed from such an innovative approach for digital voice compression based on Artificial Neural Networks (ANN) [1, 3, 6, 11], which needed a mean transmission capacity of 6.4 Kbps and generated a fairly good quality signal (according to the MOS test). The ANN compression system showed very promising results, but no comparative study with the standard algorithms used in wireless telephony was made, because the information needed to conduct the experimentation was not available at that time.

The work presented here has facilitated the execution of this type of study, by uniting theoretical and practical information about three widely used compression systems: CELP[4], VSELP[5] and GSM 06.10[2, 9]; and an improved ANN based encoder with a 2.4 Kbps capacity[10]. A comparative study of the four voice compression systems was conducted.

The theoretical background of voice compression is sketched out in section **2**. Section **3** explains the voice compression algorithms used in this study. Section **4** describes the methodology and results of the study. The conclusions are presented in section **5**.

2 Theoretical background

2.1 Human vocal tract model

The speech is produced by the vibration of the vocal chords and by the resonance of the resulting pressure wave on the vocal tract walls. In adults, the vocal tract is a tube of approximately 17 cm long with a transversal area varying between 0 and 20 cm^2 [8] (Figure 1).

The speech sounds may be classified as voiced (originated by the quasi periodical vibration of the vocal chords) or voiceless (originated by the friction of the air passing through a constriction of the vocal tract). In practice, speech sounds are a mixture of both.

During the process of generating voiced sounds, the vocal chords are kept closed, but the air pressure from the lungs forces their opening, which causes a pressure drop that closes them again, occasioning a vibration at a frequency between 50 and 400 Hz (the pitch of the sound).

The movements of the mouth, the tongue, the lips and the velum of the soft palate must be considered in addition to the vocal chords vibration and the vocal tract resonance to model the voice generation process. Therefore, a basic model of this process should be considered as follows:

- The voice is a signal emergent from a definite source: the vibration of the vocal chords for voiced sounds, and the noise caused by a constriction of the vocal tract for unvoiced ones.

- Before passing through the vocal tract, the sound wave has a relatively flat spectrum (without formants).
- The resonances of the particular shape of the vocal tract change the source spectrum, creating a formant structure that characterizes the timbre of the phoneme that is being articulated.
- If short time intervals are taken, it's possible to model the effect of the vocal tract shape as a filter which defines the relation between its input (the sound source) and its output (the generated voice).

A voice coding system that uses this model is denominated a VOCODER.

2.2 Linear Filters

The effect of the resonances of the vocal tract can be modeled by mathematical systems called *linear filters*, which are represented by their transfer function H(z). The voice is considered as a signal generated from the vocal tract filtering of a glottal quasi periodical excitation added to a random amplitude signal of uniform spectrum called *white noise* (Figure 2).

2.3 Estimation of the linear filter coefficients

Linear prediction coding (LPC) is one of the most used techniques in order to discover the transfer function of the filter that characterizes the state of the vocal tract [7]. The speech signal is analyzed by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal is called the *residue*.

The coefficients of the LPC that describe the formants and the residue can be stored or transmitted. The speech signal may be synthetized by reversing the process, using the residue as a source signal for the LPC filter.

The LPC analysis is based on the fact that the value of a signal sample $s(n)$ can be predicted from a finite number of previous samples $(s(n-1), ..., s(n-p))$, with an associated error $e(n)$, using an all-poles linear filter:

$$s(n) = e(n) + \sum_{k=1}^{p} \alpha_k s(n-k) \tag{1}$$

The prediction error (also know as residual signal), $e(n)$ is the difference between the actual signal value $s(n)$, and the prediction value $\hat{s}(n)$:

$$e(n) = s(n) - \hat{s}(n) \tag{2}$$

The coefficients α_k of the LPC filter are estimated minimizing the cuadratic error E over a window of N samples:

$$E = \frac{\sum_{i=1}^{N-1} e^2(i)}{2} \tag{3}$$

The LPC coefficients represent the envelope of the spectrum (the formant structure of the timbre) and the residual signal the source characteristics (the prosody). The transfer function of the LPC filter is:

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}} \tag{4}$$

3 Voice Compression Algorithms

In this section four voice compression algorithms will be analyzed to conduct the comparative study.

3.1 CELP

The techniques of voice compression are based on two intrinsic operations:

- redundancy reduction.
- irrelevance reduction.

The first operation uses predictions or transformations to eliminate redundant data. The second one reduces the bandwidth doing a component quantization which introduces a distortion or reconstruction error.

To increase the compression ratio, the encoder minimizes the error perception using inherent human auditory characteristics: same levels of distortion error are perceived in different ways at different frequency bands. Bands with higher energy, which correspond to formants, tolerate a higher noise level.

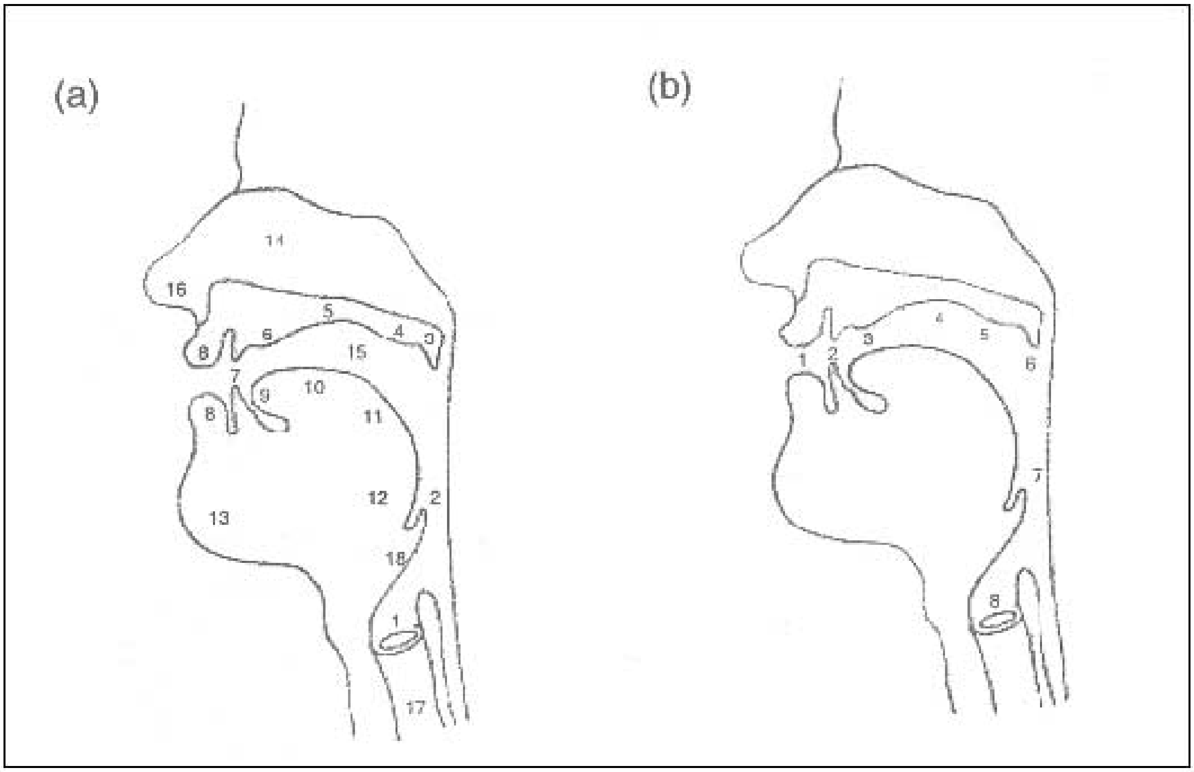

Figure 1. [7] Vocal Tract a) Speech articulations (1) vocal chords (2) pharynx; (3) veil; (4) soft palate; (5) hard palate; (6) alveoli; (7) teeth; (8) lips; (9) tongue tip; (10) tongue; (11) back; (12) root; (13) jaw; (14) nasal cavity; (15) oral cavity (16) nasal windows; (17) trachea; (18) epiglottis. b) kind of voice articulation: (1) labial; (2) dental; (3) alveolar; (4) palatal; (5) velar; (6) uvular; (7) pharyngeal; (8) glottal.

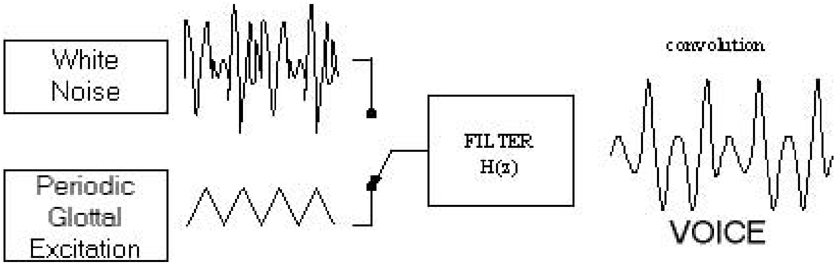

Figure 2. The voice is the convolution of a mixture of a periodic and noise excitation with the vocal tract filter response

The CELP solution [4] to compression is to use the *analysis by synthesis* approximation: A signal synthesis is made, adjusting the filter parameters and the excitation signal so that the difference between the original and the reconstructed signal is minimized.

In practice, CELP systems employ fast searching algorithms exploiting the design structure shown by the following diagram (Figure 3):

- The filter coefficients are quantified and then represented by a set of Linear Spectrum Pair (LSP).
- For pitch calculation, an autocorrelation of the original signal is made, the shorter lag showing the larger correlation is chosen, and a signal with this lag (wave length) is selected from a codebook called *adaptive codebook* (ACB).
- There is a white noise codebook (Gaussian distribution signals) called *stochastic codebook* (SCB), the best codeword with the lowest error is selected.
- A perception filter (P_W) amplifies the signal formants. In that way, an error signal with its energy concentrated at the formants position is considered better.

The decoder receives the coded parameters and rebuilds the speech signal $\hat{s}$ using the same scheme but in reverse.

3.2 VSELP

VSELP was designed by Motorola, who is responsible for the design and the development of the algorithm. It's a CELP variation that encodes to 7.950 bps, using an additional 5.050 bps for error control and frame synchronization.

The difference between VSELP and common encoders (CELP) lies mainly in the structure of its codebooks. While CELP utilizes a SCB to conduct its search, VSELP uses two vectors groups to generate a "candidate vectors" space. In that way, the search in the CELP codebook corresponds to two searches in VSELP.

There are seven orthogonal base vectors[5] in the space for each search. Each one contains 40 elements. The selection of the base vectors is fundamental for a fast search in the codebooks.

The analysis by synthesis proceeds with 3 codebooks. Firstly, by looking in the ACB for the best input and gain. This input multiplied by its gain factor is given to the first seven base vectors. Therefore, the search in the second codebook can be done independently of the first.

A new space of seven vectors is used for the second search; and a new "best" input and gain are obtained from the second codebook. Finally, a search is done in the third codebook. The gain obtained from each one of the three codebooks is quantified and transmitted with the three codebook indexes to the receptor.

The principal features of VSELP (Figure 4) are:

- Order 10 LPC analysis.
- Long term prediction.
- Adaptive codebook search (pitch estimation).
- Base vectors codebook first search.
- Base vectors codebook second search.
- vector quantization of gain.

This algorithm uses a 8 Khz sample frequency, 160 samples per frame, dividing it into 4 subframes of 40 samples.

The values β, γ_1 y γ_2 are encoded using a gain table and a "frame energy"[5]. The decoder uses the transmitted data in the same way than the standard CELP VOCODER, with the following exceptions:

- The synthesis LPC coefficients are the original (not the expanded ones).
- There is no close-loop to search the best sample.
- There is an adaptive filter for the output signal.

3.3 ANN based Encoder

This system is a VOCODER with some particularities:

1. It uses a LPC codebook instead of encoding the LPC coefficients directly.
2. The encoder and decoder have two backpropagation ANN, one for the next pitch prediction and another for the gain factor prediction.
3. The encoder sends the prediction pitch error and the gain factor error, which minimizes the transmission data size.

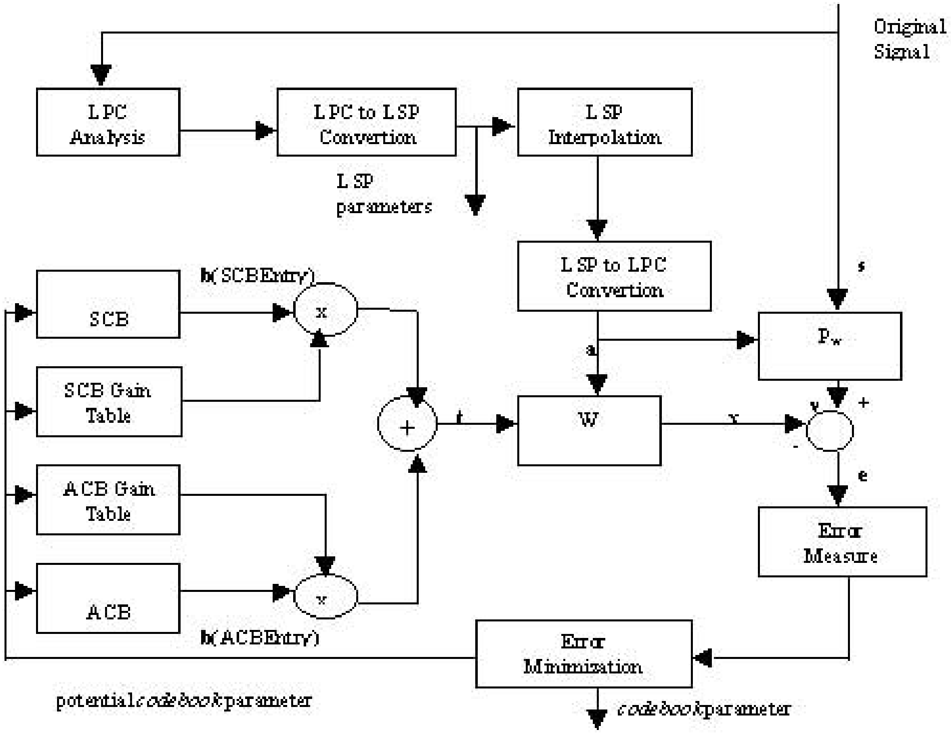

Figure 3. CELP analyzer scheme

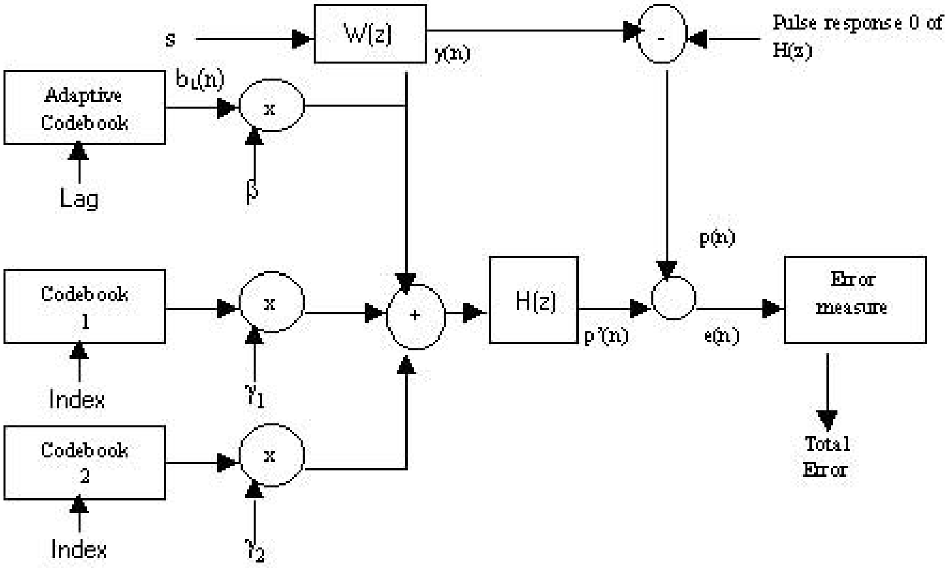

Figure 4. VSELP analyzer scheme

4. The transmission frame is pitch synchronous, not fixed size.

The LPC codebook is built from a Kohonnen's self-organizative map with a toroidal topology, which permits a very compact vector quantization of the LPC's coefficients.

3.3.1 The essential parameters

The idea of voice compression supposes that some basic forms exist whose combinations are capable of creating a complete voice signal. These basic forms are named the *essential parameters*.

The codebooks that exist in the transmitter as in the receiver, contain these basic forms of the speech (codewords). The compression consists in finding the codewords that defines the signal in the codebooks, so that it is only necessary to transmit their location. Five subsystems [11, 10] were developed to transmit the essential parameters (Figure 5):

- *Pitch detector*. Prediction using backpropagation ANN.
- *Short term LPC synthesis and search in the codebook*. The codebook is formed by the centroids of a Kohonnen's self-organizative map.
- *Long term LPC synthesis and search in the codebook*.
- *Residual signal synthesis and search in the codebook*.
- *Gain factor quantization*.

The gain factor is normally coded in 5 bits, but taking advantage of the fact that there are only slight variations from one frame to the next, the gain difference can be coded in 3 bits.

3.4 GSM 06.10 RPE-LTP

The Regular Excitation Long-Term Predictor (RPE-LTP) is the compression system used by European wireless telephony, employing the LPC technique for voice synthesis.

The GSM 06.10[2, 9] input consists of frames of 160 PCM values encoded using 13 bits. Each frame represents a 20 ms window. The encoder compresses the input of 160 values to a 260 bits frame.

The GSM 06.10 compressor models the speech with two filters and an initial excitation. The first step in the compression (and the last in the decompression) is a short term linear prediction filter that simulates the role of the vocal tract and nasal tract. It is excited by the output from a second filter of long-term linear prediction that transforms its input (the residual signal) into a mixture of glottal excitation and noise (Figure 6).

The 40 samples block of the long-term residual is represented by a sub-sequence of 13 pulses chosen among four candidates. The chosen subsequence is identified by its position in the RPE matrix (the matrix that contains the four candidates).

The algorithm chooses the sequence of maximum energy, i.e. the sequence with the highest quadratic sum of its values. An index of 2 bits transmits the selection to the decoder. This leaves 13 samples of 3 bits and a scaling factor of 6 bits (the algorithm adapts itself to the total amplitude, increasing or decreasing the scale factor).

Finally, the encoder prepares the next long-term prediction updating its "past output"; i.e., the residual short-term reconstructed signal. To ensure that the encoder and the decoder work on the same residual signal, the encoder simulates the steps of the decoder until the short-term stage.

4 A comparative study of the compression algorithms quality

The following is a small-scale comparative study with the aim of showing the relative quality of the systems of voice compression described in section 3. The study may also serve as a methodological base for future related studies.

Two *metrics* were defined to compare the algorithms: transmission capacity and signal reconstruction quality.

4.1 Transmission capacity

It refers to the minimum capacity that the transmission media should have so that all the information the encoder must send will reach the decoder in real time. It's measured in Kilobits per second.

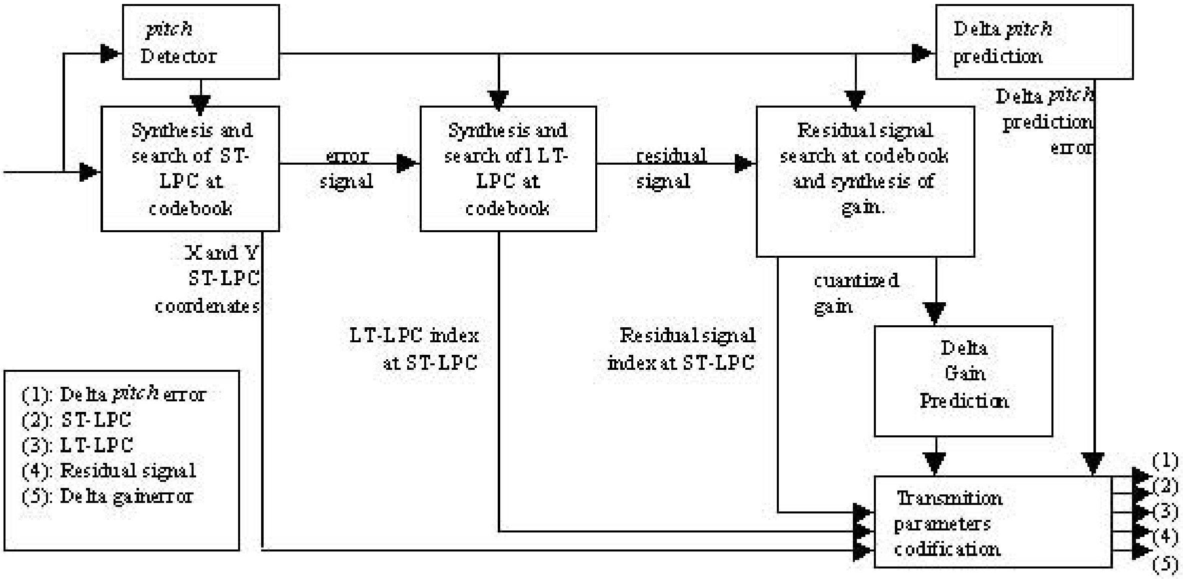
pitch Detector
Delta pitch prediction
Synthesis and search of ST-LPC at codebook
error signal
Synthesis and search of l LT-LPC at codebook
residual signal
Residual signal search at codebook and synthesis of gain.
Delta pitch prediction error
X and Y ST-LPC coordenates
LT-LPC index at ST-LPC
Residual signal index at ST-LPC
cuantized gain
Delta Gain Prediction
(1): Delta pitch error
(2): ST-LPC
(3): LT-LPC
(4): Residual signal
(5): Delta gainerror
Transmition parameters codification
(1)
(2)
(3)
(4)
(5)

Figure 5. ANN based encoder scheme

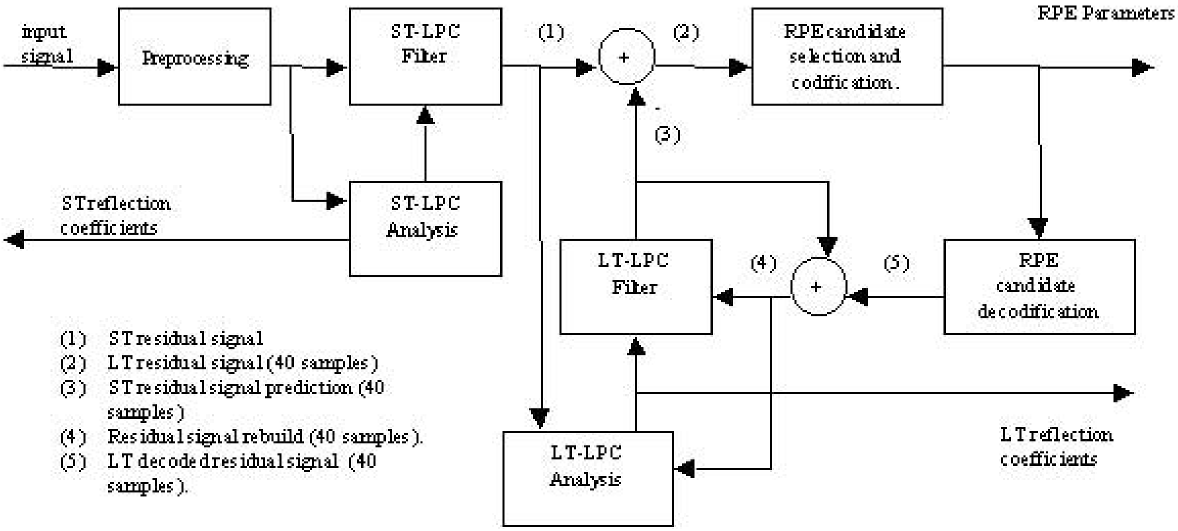
input signal
Preprocessing
ST-LPC Filter
(1)
+
(2)
-
(3)
RPE candidate selection and codification.
RPE Parameters
ST reflection coefficients
ST-LPC Analysis
LT-LPC Filter
(4)
+
(5)
RPE candidate decodification
(1) ST residual signal
(2) LT residual signal (40 samples)
(3) ST residual signal prediction (40 samples)
(4) Residual signal rebuild (40 samples).
(5) LT decoded residual signal (40 samples).
LT-LPC Analysis
LT reflection coefficients

Figure 6. GSM 06.10 encoder scheme

The following table presents the algorithms capacity:

Algorithm	ANN	CELP	VSELP	GSM
Kbps	2.4	4.3	8.0	13.0

These are theorical values: RNA was obtained from [10], CELP from [4], VSELP from [5] and GSM from [9].

4.2 MOS Test

The MOS Test was normalized by the International Consultive of Telephonic and Telegraphy committee at the 80's beginning and it has been used mainly for measuring cellular digital communication systems quality. It consists of making an opinion survey and individual tests that evaluate voice recordings according to the following tables:

VALUE	QUALITY	DEGRADATION
5	Excellent	Inaudible.
4	Good	Audible but not annoying.
3	Acceptable	Slightly annoying.
2	Mediocre	Annoying.
1	Bad	Very annoying.

VALUE	LISTENING EFFORT
5	No effort required.
4	Attention necessary, slight effort.
3	Moderate effort necessary.
2	Considerable effort necessary.
1	Whatever effort is futile for comprehension.

A web page was developed for the study with 14 samples of spanish sentences (7 from female speakers, 7 from male speakers) that were also coded and later reconstructed with the CELP, VSELP and GSM systems. It was not conducted for the ANN based encoder because a previous study [10] provides the data.

34 subjects who were not individually identified answered the test, i.e: 476 MOS evaluations were joined. The results are the average of the evaluations obtained by each sentence for male and female speakers:

	ANN	CELP	VSELP	GSM
Masculine	3.0	3.4	3.6	3.5
Feminine	3.1	3.5	3.8	3.8

To let the ANN based encoder be "comparable" with the other three algorithms, its evaluation was adjusted by an appreciation corrective factor experimentally obtained and whose value is *1.28*. To find this factor, the MOS test was repeated under the same specifications of the original study [10] using a smaller sample (84 MOS evaluations), then the resulting average was divided with those obtained previously and the average of these values was adopted as the corrective factor.

The most important point in this study is the relation between the reconstruction quality and the transmission capacity of the algorithms. This is illustrated in Figure 7.

The segmented line in the graph indicates the relation 1:1 between the average evaluation obtained by each algorithm in the MOS Test (by a scale factor of 3 for best visualization), against the transmission capacity. The importance of this relation is justified on the ground that an algorithm which achieves a large compression is useless if the quality of the signal is lost and viceversa, if the reconstruction is good but the compression poor.

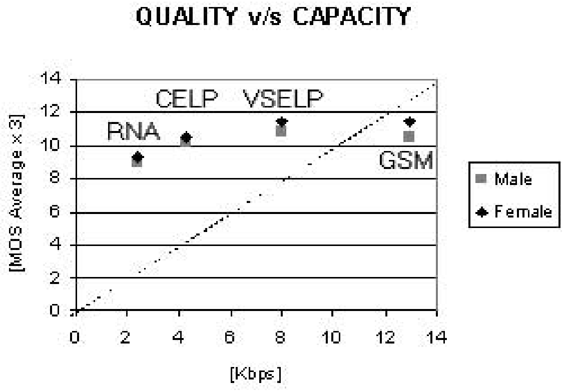

Figure 7. Graph of Quality v/s Capacity

5 Conclusions

A study of four voice compression systems has been developed, giving an explanation of the processing of the voice signal for its coding, transmission and later decoding.

As a result of the experimentation, it can be appreciated that the best performance for voice coding is the ANN based algorithm. That's because it is able to predict the parameters values in the encoder and the decoder, so that it's only necessary to send the actual and predicted differences. Furthermore, it profits from the ANN potentialities for a fast information process. However, this system cannot be used in standard telephony because it is speaker oriented: the results depend too much upon the training of the ANN (as shown in [10]).

Therefore, the general use algorithm of better performance is CELP, which is why it has been chosen as the standard of North American digital cellular telephony. This indicates that applied voice compression technology has not advanced much in more than 20 years. CELP was created in the 80's and since then the research has been dedicated to its improvement, not to the establishment of a new model that achieves a substantial improvement in the performance of voice compression for wireless telephony.

Acknowledgements

We thank the helpful comments of the referees.

References

[1] G. Cauley. "Speech production and perception". http://www.dcc.uchile.cl/∼abassi/WWW/Voz/Cauley96.ps.gz, 1996.

[2] J. Degener. "Digital speech compression: Putting the GSM 06.10 RPE-LTP algorithm to work". http://www.ddj.com/articles/1994/9412/9412b/9412b.htm, 1994.

[3] J. Freeman and D. Skapura. *"Redes Neuronales: Algoritmos, aplicaciones y tecnicas de programacion"*. Addison-Wesley Iberoamericana S.A., Wilmington, Delawere, U.S.A, 1993.

[4] G. Langi and Kinsner. "Fast celp algorithm and implementation for speech compression". http://warp.ampr-umars.umsu.umanitoba.ca/umars/research/celp1.ps, 2000.

[5] J. Macres. "Theory and implementation of the digital cellular standard voice coder: VSELP on the TMS320C5x". Application report, DSP Software Engineering, Incorporated, October 1994.

[6] D. Mondaca. "Desarrollo y simulacion de un sistema compresor de senales de voz utilizando redes neuronales". Master's thesis, Departamento de Ingenieria Electrica. Universidad de Chile, 1996.

[7] L. Rabiner and J. Biing-Hang. *"Fundamentals of Speech Recognition"*. 1993.

[8] T. Robinson. "Speech analysis". http://svr-www.eng.cam.ac.uk/∼ajr/teaching.html, 1998.

[9] E. T. Standard. "Digital cellular telecommunication system; full rate speech; transcoding (GSM 06.10 version 5.0.1)". Thecnical report, ETSI TC-SMG, May 1996.

[10] J. Velasquez. "Analisis fino del tracto vocal basado en filtros LPC aplicado al mejoramiento de la calidad de sintesis de voz". Master's thesis, Departamento de Ciencias de la Computacion. Universidad de Chile, 1996.

[11] J. Velasquez. "Aplicaciones avanzadas de redes neuronales al desarrollo y simulacion de un sistema compresor digital de voz". Master's thesis, Departamento de Ingenieria Electrica. Universidad de Chile, 1996.