# ASPECTS REGARDING SYNTHETIC SPEECH QUALITY EVALUATIONS FOR MILITARY SYSTEMS

Mihai RADU, Ştefan-Adrian TOMA and Florin POPESCU

Military Technical Academy, Communication Department
George Coşbuc Street 81st, Sector 5, Bucharest, Romania
{mradu, fdpopy}@mta.ro, stefantoma@tomas.ro

Many modern systems used by the military make use of speech processing technologies. This paper gives an overview of quality evaluations for synthetic speech generated by speech coders used in military systems. The paper is organized as follows. Section 1 is the introduction. In section 2 we make a brief presentation of speech coding technologies. Synthetic speech quality evaluation techniques and experimental results are presented in section 3. The paper ends with the conclusions.

## 1. INTRODUCTION

Because speech is the most natural mean of communication between humans, it has a very important role in military applications [6], [22]. Although an image has more information than a thousand words, in the military, there are many situations when speech offers the highest reliability, the shortest response time and the smallest workload.

Speech technologies (speech coding, speech recognition and speech synthesis) are integrated into a wide-range of military applications and systems. The key military applications areas for speech coding, as indicated in [1] are: command and control, communications, intelligence, joint forces training and operations, cockpit fast jet, helicopter.

Communications represent one of the most important fields of application for speech processing. Traditionally, this domain was linked with speech coding and speech enhancement technologies. Most of the research in speech coding for military applications is focused on low bit-rate parametric coders and delta waveform coders because they offer improved jamming resistance and a lower probability of interception. Parametric coders are based on the extraction of the vocal tract parameters. At the receiver, signal reconstruction is done through the use of rule-based speech synthesis techniques. Due to this fact we can state that the output of such a coder is a synthesized signal. Unfortunately this signal is not a perfect replica of the original input, hence quality evaluations have to be made in order to ensure that listeners would understand the message.

This paper offers an overview of quality evaluation for synthesized speech and is organized as follows. Section 1 is the introduction. In section 2 we make a brief presentation of speech coding technologies. Synthetic speech quality evaluation techniques and experimental results are presented in section 3. The paper ends with the conclusions.

## 2. CLASSES OF SPEECH CODERS

Speech is continuous-time signal and can be represented digitally through the processes of sampling and quantization. A sampled speech signal contains information which is redundant. There is considerable correlation between adjacent samples because, on the average, the signal does not change rapidly from sample to sample, so that the difference between adjacent samples has a lower variance than the variance of the signal itself. Speech coding is about finding redundancy in the signal and removing it. If only redundancy

is removed, the original signal can be recovered exactly (lossless coding). In lossy coding, the signal cannot be recovered exactly, though it sounds similar to the original.

There are three classes of speech coding techniques: waveform coding which quantizes the speech samples directly and operates at high bit rates (16-64kbps), parametric coding known as model based which quantizes the parameters of a speech production model (1.2-4.8kbps) and hybrid coding which is partly waveform based and partly model-based (2.4-16kbps). Figure 1 shows graph of speech coding quality (measured as a Mean Opinion Score (MOS)) versus bit rate for the classes of coders.

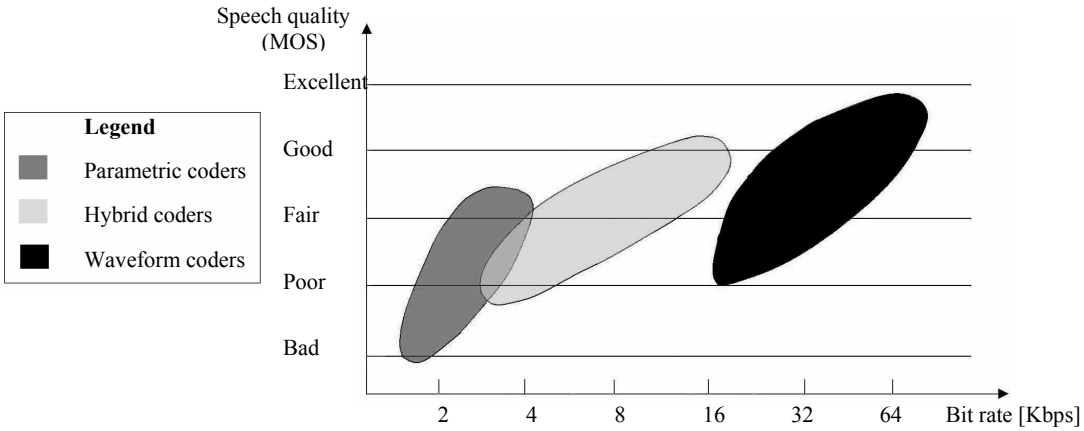


**Figure 1.** Speech quality versus bit rate for the classes of coders.

### 2.1. Waveform Coders

A waveform coder directly encodes waveforms by exploiting the temporal time domain characteristics of the speech signal. It treats speech signals as normal signal waveforms. Without using any knowledge of how the speech signal to be coded was generated, a waveform coder produces a reconstructed signal whose waveform is almost identical to the original. Waveform coders are able to produce high-quality speech at high enough bit rates. An example of a waveform coder used on large scale by NATO is the 16Kbps CVSD coding algorithm (also known as STANAG 4209). The CVSD coder is an excellent compromise between circuit simplicity and bandwidth economy. The reconstructed voice is remarkably natural, but has a slightly "fuzzy edge".

### 2.2. Parametric Coders

Increasing compression ratios means removing more redundancies in the speech. This is possible if coders exploit the physical nature of human speech and the mechanism by which it is created. Parametric coders (also known as vocoders) achieve this. This class of coders takes advantage of the high redundancy /correlation properties of speech by modelling the human vocal tract and extracting parameters that are necessary in generating speech waveforms. In vocoders, not the signal samples but the parameters of a source-filter speech model are quantized and transmitted. Vocoders produce intelligible speech at very low bit rates, but they sometimes sound mechanical or buzzy and are prone to annoying thumps and tonal noises. There are two parametric coders used in NATO: the LPC-10 (STANAG 4198) and MELP (STANAG 4591), both at 2.4Kbps. The MELP achieved a MOS score of 3.3 with a quiet background compared with 2.2 for LPC-10 [2].

### 2.3. Hybrid Coders

The speech quality of waveform coders drops rapidly for bit rates below 16 kpbs, whereas there is a negligible improvement in the quality of vocoders at rates above 4 kpbs (figure 1). Hybrid coders are thus

used to bridge this gap, providing good speech quality at medium bit rates. However, these coders tend to be more computationally demanding. Virtually all hybrid coders rely on LPC analysis to obtain the synthesis model parameters. Waveform coding techniques are then used to code the excitation signal. Pitch production models can be incorporated to improve the performance [5]. An example of such a coder is the 4.8Kbps CELP (U.S. Federal Standard 1016) used by the U.S. government.

More information regarding speech coding for military communications can be found in [4], [5], [11] and [21].

## 3. SPEECH QUALITY EVALUATION

Evaluation is the process by which an object, algorithm, situation, etc is assessed based on appropriate criteria and indicators. Based on the indicators one can make decisions regarding the object, algorithm or situation under evaluation.

In military communications the evaluation of speech coding techniques has an important role both in the design phase and during operations. In the design phase evaluations are made with regard to the number of necessary computations, jamming resistance, speech quality, weight, and cost. During military operations only real-time speech quality evaluations are made but based on the results some of the other parameters of the algorithm can change.

The quality of a synthesized speech utterance is its property of being natural and intelligible. When we speak of intelligibility we refer to the property of the synthesized speech utterance of transmitting the information imbedded inside it. For example let us consider a situation where in a noisy environment the user at the emitter side of the connection transmits the utterance “Launch at 02.00”. Because of the environment the user at the receiver end understands the message but he is unable to identify the person transmitting (even if it is someone he would normally recognize). This means that the transmission is only intelligible.

Naturalness is the property of the synthesized speech utterance of being perceived as close as possible to the human generated speech.

Thus, the evaluation process can be made on the speech quality, speech intelligibility or speech naturalness.

The evaluation process can be classified based on four criteria: the analysis instruments, the environment where the evaluation is conducted, transparency and the analyzed elements [3]. From the point of view of the analysis instruments criterion the evaluation process can be subjective when it is made with human operators or objective when the evaluation instruments are computers or other digital signal processors. The second criterion, the environment, refers to whether the evaluation is conducted in laboratory conditions or online. The transparency criterion refers to whether during the evaluation the algorithm is considered as a “black box” or as a system composed of many elements. Finally, the criterion analyzed components refers to whether the evaluation is made globally or only some aspects are analyzed.

In the following we will discuss the problem of subjective and objective speech quality and intelligibility.

### 3.1. Speech Quality

From a historic perspective the first methods for evaluating the quality of a synthesized speech utterance or a transmission were based on signal to noise ratio (SNR), harmonic distortion or bit error rate (BER) based measurements [7,15,18]. With the advent of digital speech coding techniques (delta modulators, LPC, perceptual coding) these type of measurements became obsolete due to the fact that by the use of the new algorithms one can generate a low SNR speech utterance when compared with the original speech utterance, that is perceived by a human user as a good quality and intelligible speech. To solve this problem, new techniques for speech quality evaluations, that take into account the perception properties of the human year, were developed.

*Subjective speech quality evaluation*

Subjective speech quality evaluations have been used for a long time, at first for intelligibility measurements and then for quality evaluations. Methods for subjective quality evaluations were standardized by ITU-T as Recommendation P.800. Two type of test are described in the recommendation: conversation-opinion tests and listening-opinion tests. Conversation opinion tests are conducted in laboratory and are intended to reproduce the actual service conditions experienced by the user of the system. Listening-opinion tests are also laboratory but here users listen to speech utterances – short phrases 2 to 5 seconds long, and then grade them on some scale [7]. The types of listening test presented in Recommendation P.800 are: Absolute Category Rating (ACR), Degradation Category Rating (DCR) Comparison Category Rating (CCR), Quantal-Response Detectability Tests (QRDT) and the threshold method for comparison of transmission systems with a reference system. The ACR test represents an absolute evaluation of the system whit three types of results depending on the scale that was used: Mean Opinion Score (MOS), Mean Opinion Score Listening Effort ($MOS_{LE}$) and MOS Loudness Preference ($MOS_{LP}$) [7]. DCR represents an alternative to ACR. In this type of test the operator grades a synthesized speech utterance when compared to the original signal on a scale ranging from „Inaudible" to „Very annoying". The result is called Degradation Mean Opinion Score (DMOS) [7]. CCR is a variation of the DCR where the scale ranges from „Much better" to „Much worse" [7]. For the speech algorithms presented in the first part of the paper MOS and DMOS evaluations are presented in figures 2 and 3.

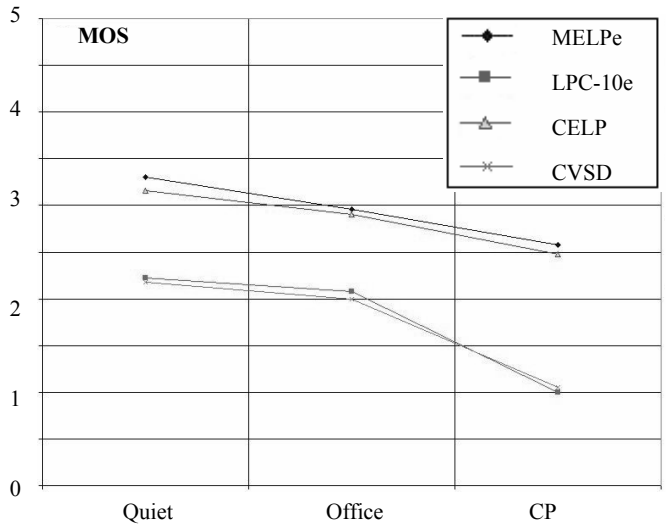


**Figure 2.** MOS versus environment for speech coders used by NATO and the USA (CP – Mobile Command Point) [12].

In figure 2, Mean Opinion Score results for four speech coders in increasing noisy environments are presented. One can clearly observe the decreasing in quality with the increase in noise.

*Objective speech quality evaluation*

Subjective speech evaluation techniques produce valid scores but unfortunately most of the times they are impractical: to assemble a large number of listeners is both difficult and expensive. Thus, new, objective techniques had to be developed. As we have stated before in our paper the first evaluation techniques were based on SNR or BER measurements [7], [15, [18]. An example of a SNR speech evaluation technique uses the segmental SNR (S-SNR). The S-SNR calculates the SNR on short intervals and then averages the individual measures [15].

Because SNR and BER evaluation methods do not use psychoacoustic information their results cannot be applied to modern coding systems. In recent years modern objective speech evaluation tests were developed to cope with the new requirements. They can be classified in two categories [10]:

- Signal based methods
- Parameters based methods.

Signal based methods use the original signal as a reference input for the measurement and then compares it with the signal to be evaluated (its degraded version) [20]. The comparison is made by the use of spectral distances consistent with the physiology and psychology of the hearing process. Examples of such spectral distances are the Log Spectral Distance, Mel Cepstral Distance, Bark Scale Distance etc. We have evaluated the speech quality generated with three coders (MELPe, CELP and LPC-10e) by using the Log Spectral Distance. The results are presented in figure 4.

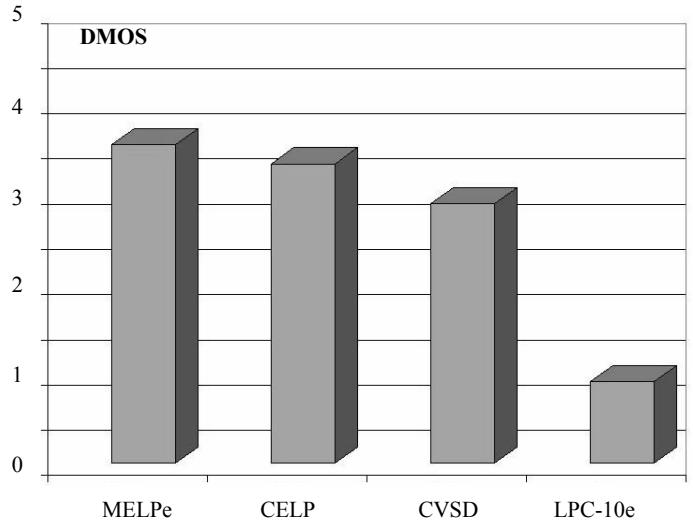


**Figure 3**. DMOS for speech coders used by NATO and the USA [12]

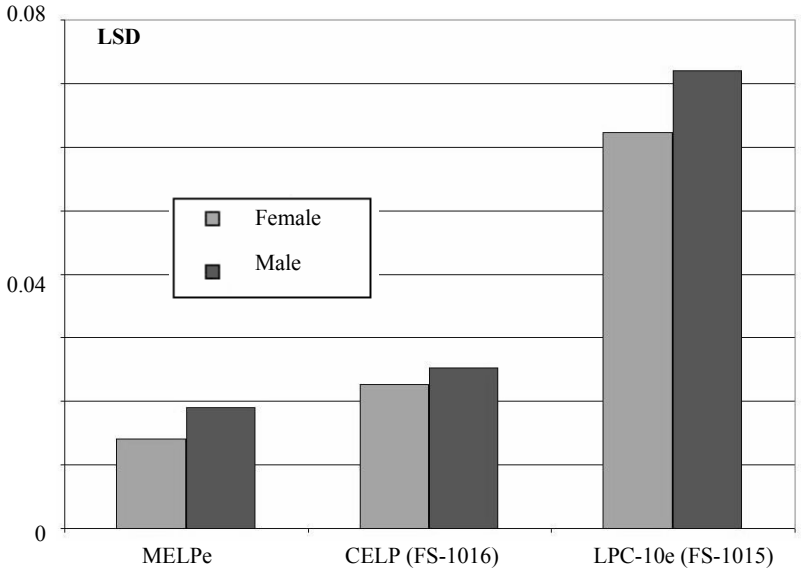


**Figure 4.** Log spectral distances for speech coders used by NATO and the USA government.

The results in figure 4 are for both male and female speakers and can be correlated with the DMOS results: in both graphs the MELP and CELP coders cluster together with the LPC10 coder having a much lower quality (small values indicate high quality).

There are standardized signal reference techniques like the Perceptual Speech Quality Measure developed in the 1990s by KPN Research in the Netherlands and adopted by ITU-T as Recommendation P.861.

Although a very important tool for objective speech evaluation PSQM doesn't include time alignment and level alignment. To fix these problems a more capable algorithm was developed - Perceptual Estimation of Speech Quality, adopted by ITU-T in February 2001 as Recommendation P.862. Further information regarding the PESQ algorithm can be found in [8].

PESQ is probably the most used speech evaluation tool nowadays but one must be advised to consider the problem of correlation with subjective results: although the average correlation with subjective evaluations is 0.935 different results are obtained for different languages [16].

Signal reference methods are also called intrusive methods because the system needs to be taken out of service during the evaluation process.

The subjective evaluation techniques and some of the objective signal reference evaluation techniques offer reliable results but are unsuitable for real time monitoring of speech communications. Unlike those, the parameter based techniques estimate the speech quality based on computational models, making the need for a reference signal obsolete. Thus, the parameters based methods are suitable for real time monitoring of speech communications. Because of this they are also called non-intrusive evaluation methods.

One such method proposed in [10] uses the digital water marking technique for estimating the quality of speech transmissions. The digital watermark is embedded in the discrete wavelet transform of the signal. When passing through the system, the watermark is degraded along with the signal. At the receiver, the quality is estimated by calculating the Percentage of Correctly Extracted Watermark bits and by mapping it with the PESQ MOS scores [10]. A detailed description of how the watermark is embedded in the signal and how the mapping between PCEW and MOS is made can be found in [10].

Based on this technique we propose a system that makes real-time speech quality adjustments by modifying several parameters (see figure 5).

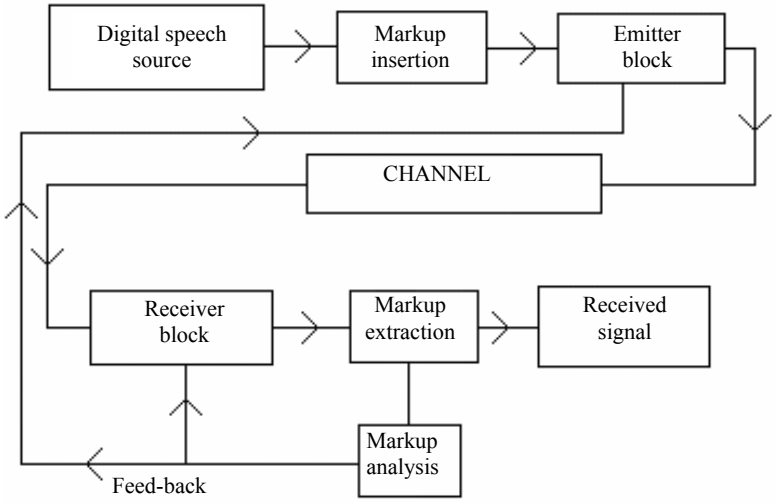


**Figure 5.** Communication system with automatic quality adjustment.

It can be observed from figure 5 that after the analysis of the markup, decisions regarding the quality of the speech are made and commands for the emitter block and receiver block are sent through a feed-back channel. The advantage of such a system is that it can detect changes in the quality of the received signal before they became irritating for the user. At the moment we are building a Matlab simulation for this system.

### 3.2. Speech Intelligibility

For a correct evaluation of a speech coder quality tests are insufficient. Due to the fact that the quality is a two component property, it is possible for a coder to score low in a DCR test for a harsh environment situation but score high with an intelligibility test in the same environment. For example, in [12] are

presented evaluation results for four coders used by the US government: LPC-10e (FS-1015), CELP (FS-1016), MELPe and CVSD (STANAG 4209). The CVSD coder scores almost a point lower on the DMOS scale, when compared with the MELP coder, in an airplane environment, yet it scores higher than the MELP coder in a Dynamic Rhyme Test for the same conditions. This happens because the degradation category test measures both intelligibility and naturalness.

Intelligibility measures are based on the response of human subjects to some speech utterances. The result of an intelligibility test is the ratio between the number of correctly identified speech utterances and the number of transmitted ones.

Two of the most used tests nowadays are the Dynamic Rhyme Test and the Modified Rhyme Test [9], [13], [14], [15], [17], [19].

Articulation tests with *logatoms* are a version of the MRT tests. Such tests were carried out at the Military Technical Academy, for the first time in our country, for several types of delta coders with plans to use the same test for the evaluation of the LPC10 and MELPe coders. In the following we concisely present the methodology and the results of the evaluations.

A logatoms is a phonetic segment similar to a syllable but without meaning. Several operators are used in an intelligibility test with logatoms. One of the operators reads a logatom table, while the rest of them listen - in another room - to the transmission (see figure 6).

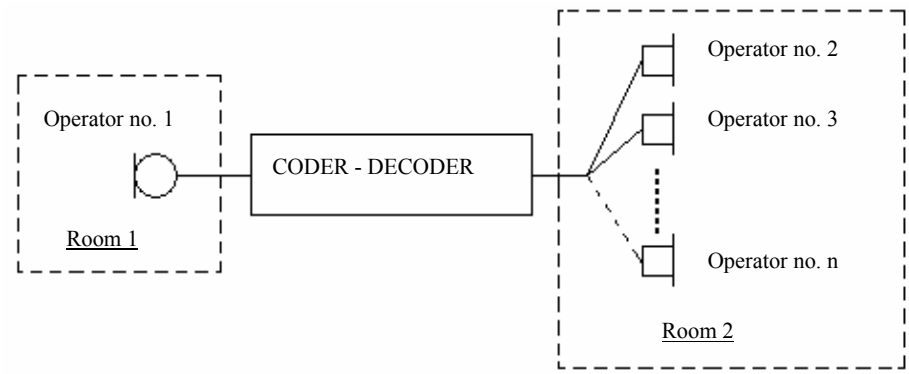


**Figure 6.** Block diagram for a logatom intelligibility testing system [18].

An example of a logatom table for the Romanian language is presented in table 1.

**Table 1.** Logatoms [18]

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Z<br>NES | CRU | PÎM | DIL | G<br>MEC | ŞAM | G<br>VIC | DAZ | G<br>TAC | D<br>DET |
| D<br>LAT | ŢAN | BUL | NEA | GÎR | FLI | Ş<br>LOJ | PRU | ŞIH | PRE |
| SEI | CUL | TRU | CISC | LEA | TUL | PIA | CESI | MEU | STĂ |
| RĂL | VEA | PER | NEP | NIR | NOA | CRI | ROA | SCO | PIE |
| DINS | FRON | D<br>CUNT | MOAR | TUŞI | D<br>TANT | CLEH | STRĂ | T<br>TIUD | BORI |

A logatom is considered correctly received if the sounds represent the main logatom variant or to the secondary variant (the one on top in some of the cells in table 1).

The intelligibility for one table ( $S_K$ ) with logatom is calculated as follows [18]:

$$S_K = \frac{S_C}{S_O} \cdot 100 \tag{1}$$

where $S_C$ is the number of correctly received logatoms, $S_0$ is the number of logatoms in the table and $K$ the table number. The average intelligibility ( $\overline{S}$ ) for $n$ tables is [18]:

$$\overline{S} = \frac{1}{n}\sum_{K=1}^{n} S_K \tag{2}$$

If the number of $n$ of tests is high enough, $\overline{S}$ will converge to the real value, $S$, of the intelligibility. Thus, an important aspect is the minimum number of tables needed to make a correct evaluation. The relation between the minimum number of logatom tables and the error of the measurement is presented in [18].

The test conditions were the following [18]:

- the rooms were insulated from noise;
- a metronome was used to keep the same rate of logatom transmission (20 logatoms/minute);
- the logatoms were pronounced with no accents, repetitions or lengthening.

The operators were chosen with respect to the following aspects [18]:

- the minimum number of operators is 7;
- the operators were from the same age group - 18 to 30 years old;
- the operators were high school graduates.

Intelligibility evaluations with logatoms are made in two phases: a training phase and an evaluation phase. The purpose of the training phase is to exclude the subjective factors from the evaluation. The training is considered complete when the average intelligibility deviation is smaller than 3% [18].

In figure 7 are presented the tests results for three types of delta modulators.

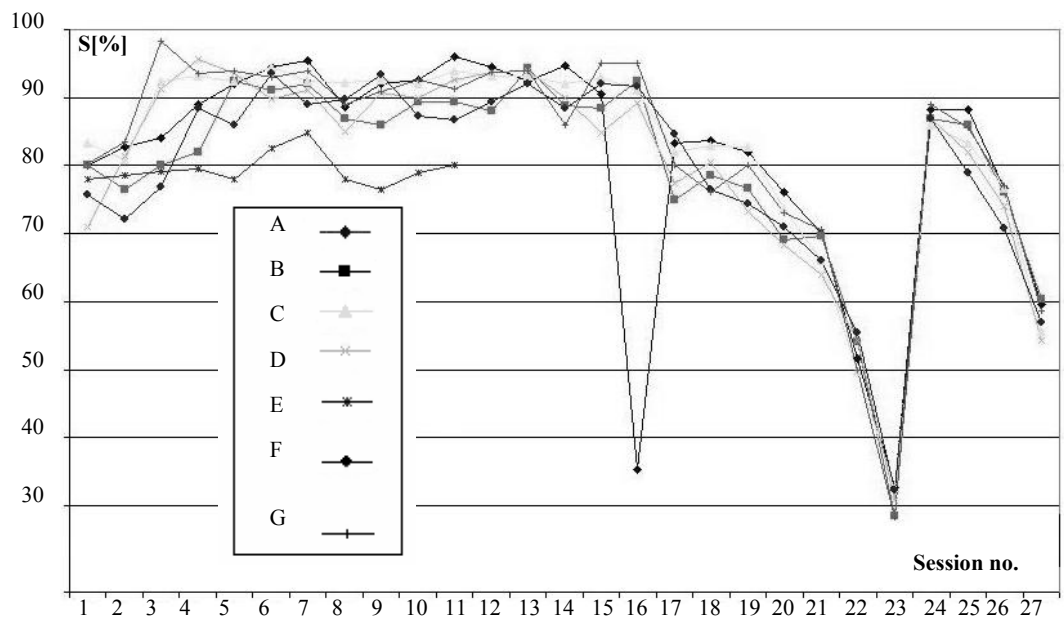


**Figure 7.** Articulation test for three delta modulators (see Table 1 for details) [18].

**Table 2.** Evaluation sessions

| Session no. | 1-13 | 14-16 | 17 | 18,19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Line speed [Kbps] | 51 | 25.6 | 16 | 16 | 14 | 12 | 128 | 64 | 150 | 150 | 150 | 150 |
| CODEC | Delta modulator with syllabic adaptation | | | | | | Linear delta modulator | | | | | |
| Speech signal input level [dB] | -15 | | | | | | | | | -28 | -45 | -50 |

The tests were conducted with a team composed of 6 listeners. Initially there were 7 listeners but during the training period – sessions 1 to 12 – one of the listeners (line E in figure 7) was not consistent with the rest of the team. A list with logatoms for the Romanian language was used for the training. After the training the team reached an average precision of 3%.

The coders were evaluated in the following way:

- the 51.2 Kbps delta modulator with syllabic adaptation (D.M.S.A.) was evaluated in sessions 13, 16 and 17 for 51.2 kbps, 25.6 Kbps and respectively 16 Kbps. The following intelligibility values were obtained: 93.33%, 92.43% and respectively 80.38%;
- the 16 Kbps delta modulator with syllabic adaptation was evaluated in sessions 19, 20 and 21 for 16 Kbps, 14 Kbps and respectively 12 Kbps. The following intelligibility values were obtained: 78.16%, 71.84% and respectively 67.72%;
- during sessions 19, 20 and 21 a 150 Kbps linear delta modulator was evaluated for 128 Kbps, 64 Kbps and respectively 150 Kbps. The following intelligibility values were obtained: 52.48%, 30.84% and respectively 87.52%.

The results are summarized in table 3.

**Table 3.** Evaluation results

| Coder type | Inteligibility [%] |
|---|---|
| Delta Modulator with Sylabic Adaptation 1 51.2Kbps | 93,33 |
| Delta Modulator with Sylabic Adaptation 1 25.6Kbps | 92,43 |
| Delta Modulator with Sylabic Adaptation 1 16Kbps | 80,38 |
| Delta Modulator with Sylabic Adaptation 2 16Kbps | 78,16 |
| Delta Modulator with Sylabic Adaptation 2 14Kbps | 71,84 |
| Delta Modulator with Sylabic Adaptation 2 12Kbps | 67,72 |
| Linear Delta Modulator 64Kbps | 30,84 |

The intelligibility is considered to be very good for $S \geq 85\%$, good for $85\% > S \geq 75\%$, satisfactory for $75\% > S \geq 70\%$ and unsatisfactory for $S < 70\%$.

## 4. CONCLUSIONS

Synthesized speech quality evaluation methods are indispensable tools both in the design phase and during operations. The evaluation process can be categorized as follows: subjective or manual quality evaluation and objective or automated quality evaluation. There are two types of objective quality evaluations: intrusive (non real-time) and non-intrusive (real-time) evaluations. At present, objective quality evaluation techniques have reached a relatively mature stage of development. There are state of the art evaluation methods (e.g. PSQE) which produce results highly correlated with the results obtained through subjective tests. Although not used as much as the intrusive methods, real-time quality evaluation techniques represent a promising direction in the field of synthetic speech quality evaluations.

Another important aspect is that speech quality must be evaluated based on several different tests, thus one will have a multi-dimensional perspective on the system that generated the synthetic speech.

With the advent of algorithms which produce high quality synthetic speech used by the military in command and control systems one will have to take into consideration the naturalness criterion besides the quality and intelligibility criteria for synthetic speech evaluations. At the moment there is no ITU-T recommendation for naturalness evaluations or any other kind of established method. We conclude that further work needs to be done in this direction.

For the first time, delta coders were tested with logatoms for the Romanian language. The logatoms used in the tests were developed by a team of researchers from the army's research institute. The results confirmed the quality of the delta coders, evaluated through the articulation tests.

Although the use of logatoms makes the test more difficult because the evaluation team of operators has to be trained, the results are more rigorous due to the fact that there is no redundancy in the speech utterance.

## REFERENCES


1. AC/243 (Panel 3) TR/21, *Potentials of Speech and Language Technology Systems for Military Use: an Application and Technology Oriented Study*, 1996.
2. GOLDBERG R., RIEK L., *A Practical Handbook of Speech Coders*, CRC Press, 2000.
3. McCREE A., BARNWELL T., *A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding*, IEEE Transactions on Speech and Audio Processing, Vol. 3, No. 4, July 1995, pp. 242-250.
4. NATO C3 Agency, *Host laboratory role in the selection of the future NATO narrow band voice coder*.
5. PEREIRA W., *Modifying LPC Parameter Dynamics to Improve Speech Coder Efficiency*, M.Eng. Thesis, McGill University Montreal, 2001.
6. *Use of speech and language technology in military environments*, NATO RTO Technical Report, Published December 2005.
7. Rec. ITU-T P.800.
8. Rec. ITU-T P.861.
9. HUANG X., ACERO A., HSIAO-WUEN HON, *Spoken Language Processing – A guide to Theory, Algorithm and System Development*, Prentice Hall, 2001.
10. LIBIN CAI, JYING ZHAO, *Evaluation of speech Quality Using Digital Watermarking*, IEICE Electronics Express, **Vol. 1**, No. 13, Pg. 380-385.
11. SPANIAS A.S., *Speech Coding: A Tutorial Review*, Proceedings of the IEEE, **Vol. 82**, No. 10, October 1994.
12. KOHLER M., *A Comparison of the New 2400 Bps MELP Federal Standard with Other Standard Coders*, p. 1587, 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97) – **Vol. 2**, 1997.
13. HOUSE A., *Articulation Testing Methods: Consonantal Differentiation with a Close Response Set*, Journal of the Acoustical Society of America, 1965, 37, pp. 293-296.
14. HOWARD-JONES P., *SOAP Speech Output Assessment Package*, ESPRIT SAM-UCL-042, 1996.
15. HOLMES J., HOLMES W., *Speech Synthesis and Recognition 2nd Edition*, Taylor & Francis, London and New York, 2001.
16. RIX A., W., *Comparison between subjective listening quality and P.862 PESQ score*, Online Workshop MESAQIN 2003.
17. NYE P., GAINTENBY J., *The Intelligibility of Synthetic Monosyllabic Words in Short, Syntactically Normal Sentences*, Haskings Laboratories, 1974.
18. RADU M., *Contribuţii la studiul sistemelor cu modulaţie delta,* Doctoral dissertation, Military Technical Academy, Bucharest, 1984.
19. VOIERS W., SHARPLEY A., HEHMSOTH C., *Research on Diagnostic Evaluation of Speech Intelligibility*, Air Force Cambridge Research Laboratories, Bedford, MA, 1975.
20. RIX A.W., BEERENDS J.G., HOLLIER M.P., and HEKSTRA A.P., *Perceptual evaluation of speech quality – a new method for speech quality assessment of telephone networks and codecs*, Proc. IEEE ICASSP pg. 749-752, May 2001.
21. POPESCU F., TOMA S. A., *Speech Coding in Military Communication Networks*, New Challenges in the Field of Military Sciences, Budapest, Hungary, 2005.
22. *Speech coding*, Lecture Notes, Data Compression and Multimedia Communication (DCMC) Laboratory, Department of Electrical Engineering, National Cheng Kung University.