ASPECTS REGARDING SYNTHETIC SPEECH QUALITY EVALUATIONS FOR MILITARY SYSTEMS

Gl. (r) prof dr. ing. Radu Mihai
Lt. drd. ing. Toma Ştefan-Adrian
Mr. drd. ing. Popescu Florin
1. Introduction
2. Classes of speech coders
3. Speech quality evaluation
   • the concept of speech quality
   • subjective evaluation
   • objective evaluation
   • intelligibility evaluation
4. The conclusions
Introduction

1. Speech technologies are integrated into a wide-range of military applications and systems.

2. Most of the research in speech coding for military applications is focused on low bit-rate parametric coders and delta waveform coders.

3. The output of a parametric codec is a synthesized voice signal.

4. Quality evaluations have to be made in order to ensure that listeners would understand the message.
**Classes of speech coders**

**Waveform coders**

Waveform coders directly encode waveforms by exploiting the temporal time domain characteristics of the speech signal. An example of a waveform coder used on large scale by NATO is the 16Kbps CVSD coding algorithm (also known as STANAG 4209).

**Parametric coders**

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. 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.

**Hybrid coders**

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. An example of such a coder is the 4.8Kbps CELP (U.S. Federal Standard 1016) used by the U.S. government.
Classes of speech coders

![Diagram showing speech quality versus bit rate for different classes of coders.]

Legend:
Blue: Parametric Coders
Yellow: Hybrid Coders
Red: Waveform coders

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

**The concept of speech quality**
The quality of a synthesized speech utterance is its property of being **natural** and **intelligible**.

**The intelligibility** is the property of the synthesized speech utterance of transmitting the information imbedded inside it.

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

**Classification of the evaluation processes**
The evaluation process can be classified based on four criteria:

- the analysis instruments (**objective** or **subjective**)
- the environment where the evaluation is conducted
- transparency
- analyzed elements.
Subjective speech quality evaluation

Subjective quality evaluations methods were standardized by ITU-T as Recommendation P.800.

Two type of test are described in the recommendation:
- conversation-opinion tests
- listening-opinion tests.

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)
- the threshold method for comparison of transmission systems with a reference system.
Subjective speech quality evaluation

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

Figure 3. DMOS for speech coders used by NATO and the USA
Objective speech quality evaluation methods can be classified in two categories:

- signal reference based methods
- parameters based methods

**Signal reference based method example**

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

- 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.
- PSQM doesn’t include time alignment and level alignment.
- Perceptual Estimation of Speech Quality, adopted by ITU-T in February 2001 as Recommendation P.862.

- Signal reference methods are 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.
- The parameter based techniques estimate the speech quality based on computational models, making the need for a reference signal obsolete.
- Parameters based methods are also called non-intrusive evaluation methods.
Objective speech quality evaluation

One such method proposed in *Libin Cai* and *Jiying Zhao* 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. The quality is estimated by calculating the Percentage of Correctly Extracted Watermark bits and by mapping it with the PESQ MOS scores.

**Figure 5.** Communication system with automatic quality adjustment (At the moment we are building a Matlab test-bed for this system)
Speech intelligibility

- For a correct evaluation of a speech coder quality tests are insufficient.
- 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.
- 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.
Speech intelligibility

- 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 logatom is a phonetic segment similar to a syllable but without meaning.

Figure 6. Block diagram for the logatom intelligibility testing system
Speech intelligibility

The intelligibility for one table with logatom is calculated as follows:

\[ S_K = \frac{S_C}{S_O} \cdot 100 \]

The average intelligibility for \( n \) tables is:

\[ \bar{S} = \frac{1}{n} \sum_{K=1}^{n} S_K \]

If the number of \( n \) of tests is high enough, 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.
Speech intelligibility

The test conditions were the following:

- 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:

- 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
- an evaluation phase

<table>
<thead>
<tr>
<th>Session no.</th>
<th>13</th>
<th>14</th>
<th>16</th>
<th>17</th>
<th>18,19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>26</th>
<th>27</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line speed [Kbps]</td>
<td>51</td>
<td>25.6</td>
<td>16</td>
<td>16</td>
<td>14</td>
<td>12</td>
<td>128</td>
<td>64</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td></td>
</tr>
<tr>
<td>CODEC</td>
<td>Delta modulator with syllabic adaptation</td>
<td>Linear delta modulator</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speech signal input level [dB]</td>
<td>-15</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Speech intelligibility – Evaluation results

-5 0-4 5-2 8-1 5

Linear delta modulator
Delta modulator with syllabic adaptation
CODEC

<table>
<thead>
<tr>
<th>Session no.</th>
<th>4</th>
<th>B</th>
<th>14</th>
<th>16</th>
<th>17</th>
<th>18,19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>26</th>
<th>27</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line speed  [Kbps]</td>
<td>51</td>
<td>25.6</td>
<td>16</td>
<td>16</td>
<td>14</td>
<td>12</td>
<td>128</td>
<td>64</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CODEC</td>
<td>Delta modulator with syllabic adaptation</td>
<td>Linear delta modulator</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Speech intelligibility – Evaluation results

<table>
<thead>
<tr>
<th>Coder type</th>
<th>Intelligibility [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delta Modulator with Sylabic Adaptation 1 51.2Kbps</td>
<td>93,33</td>
</tr>
<tr>
<td>Delta Modulator with Sylabic Adaptation 1 25.6Kbps</td>
<td>92,43</td>
</tr>
<tr>
<td>Delta Modulator with Sylabic Adaptation 1 16Kbps</td>
<td>80,38</td>
</tr>
<tr>
<td>Delta Modulator with Sylabic Adaptation 2 16Kbps</td>
<td>78,16</td>
</tr>
<tr>
<td>Delta Modulator with Sylabic Adaptation 2 14Kbps</td>
<td>71,84</td>
</tr>
<tr>
<td>Delta Modulator with Sylabic Adaptation 2 12Kbps</td>
<td>67,72</td>
</tr>
<tr>
<td>Linear Delta Modulator 64Kbps</td>
<td>30,84</td>
</tr>
</tbody>
</table>
The conclusions

• There are state of the art objective 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.
• 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.

• 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.
• 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.
The conclusions

• 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.
Thank you!