Interoperable Voice Communications: Test and Selection of STANAG 4591

M. D. Street
Communications Systems Division
NATO C3 Agency
PO Box 174
2501CD, The Hague, The Netherlands
E-mail: Michael.Street@nc3a.nato.int

J. S. Collura
Office of INFOSEC Research and Engineering
National Security Agency
9800 Savage Road, STE 6516
Ft. George G. Meade, MD 20755-6516, USA
E-mail: jscollu@alpha.ncsc.mil

ABSTRACT

This paper describes the evaluation, test and selection of STANAG 4591, the future NATO Narrow Band Voice Coder (NBVC). STANAG 4591 will enable high performance interoperable secure speech communications in harsh acoustic and channel error environments at rates of 1.2kbps and 2.4kbps. The low data rate requirements of this voice coder allow it to operate on channels with low throughput, such as wireless mobile networks. However, applying low rate voice coding to channels carrying mixed voice and data traffic allows increased capacity for data. This voice coder will enhance the quality and intelligibility of voice communications to NATO commanders in the field.

This paper describes the process to select STANAG 4591, the future NATO Narrow Band Voice Coder. The selection process has tested voice coders in a wide range of representative noise environments and conditions. Voice coders being tested have been submitted by NATO member nations. Of these, the voice coder, which offers the highest combined performance, will be adopted as STANAG 4591. The Ad Hoc Working Group is scheduled to make a preliminary selection of STANAG 4591 in late October 2001.

INTRODUCTION

The performance of current NATO voice coding algorithms – STANAG 4198 (2.4Kbps LPC10e) and STANAG 4209 (16Kbps CVSD) on narrow band channels no longer represents the state-of-the-art in voice coding technology. Neither LPC10e nor CVSD provide high quality speech in benign conditions, and their performance is degraded substantially when used in the harsh acoustic environments that are typical to current NATO operations (e.g. helicopters, tracked vehicles, supersonic aircraft).

Technological advances such as powerful noise pre-processing algorithms now allow the development of voice coders with improved performance. Some of these improvements include...

---

1Michael Street is the NC3A delegate to the NATO Ad-Hoc Working Group on Narrow Band Voice Coding AC322(SC/6-AHWG/3).
2John Collura chairs the NATO Ad-Hoc Working Group on Narrow Band Voice Coding.
superior performance as measured by speech intelligibility, speech quality and speaker recognizability. Other advances extend this improved performance in both noisy acoustic and channel error conditions. Finally, the three candidate coders all exhibit superior non-native speaker intelligibility.

Advances in digital signal processing hardware allow practical implementations of high performance voice coding algorithms in user terminals. However, the standardisation on one of these new voice coders is essential if the vital element of interoperability is to be added to the technical benefits available.

VOCODER TEST AND SELECTION PROCESS

NATO nations were invited to nominate extant voice coders for adoption by NATO. The primary requirement on the candidates was that they must operate at two bit rates (1.2Kbps and 2.4kbps) based upon the same core algorithm (modifying the quantization methods to achieve the alternate bit rates). Furthermore, that the end to end delay imposed by the voice coder must not exceed 250 ms (including a back-to-back channel). To select the optimum Narrow Band Voice Coder (NBVC) for use within NATO, an Ad Hoc Working Group (AHWG) was established – AC322(SC/6-AHWG/3). This group devised the selection process and test plan, which are currently being carried out, with NC3A as host nation.

For reasons of practicality, the test plan to select STANAG 4591 has three stages, described in detail in later sections. Phase I was a coarse test of candidate vocoders conducted using floating point C source code. Phase II is a more exhaustive test, which required that the candidate nations submit their coders implemented in fixed-point (ETSI) C source code. The preliminary selection of the coder is made using the results of Phase II. Upon the preliminary selection of STANAG 4591, the Phase III validation process is initiated. Phase III testing is primarily a real-time communicability evaluation of the selected voice-coding algorithm. Phases I and II were conducted at the NC3A-NL (for the processing of speech through the coders) and three national speech processing laboratories (for the analysis of the processed speech). Phase III will be conducted by a national test laboratory under the direction of AHWG NBVC and NC3A.

To provide the NATO member nations with a direct comparison between the candidate coders and known references deployed throughout the tactical infrastructure of the NATO member nations, three reference coders were included. These coders were the 16Kbps CVSD algorithm (STANAG 4209), the 4.8Kbps CELP algorithm (U.S. Federal Standard 1016) and the 2.4Kbps LPC10e algorithm (STANAG 4198).

PHASE I TESTING

Phase I of the selection process assessed the performance of the candidate voice coders for both the intelligibility and voice quality in benign acoustic conditions. Non real-time floating point C software implementations of candidate vocoders were used. The candidates operated on digital speech recorded as standard raw audio files (audio sampled at 8kHz, 16 bits per sample, and 2’s complement arithmetic). The voice encoders read in a raw audio file and produced a packed digital bit stream (written to file) at either 1.2Kbps or 2.4Kbps. The candidate decoders read the packed digital bit stream and generated corresponding synthetic output speech (also written to file, in raw audio format).

The voice coders were installed on NC3A’s ‘speech processing workstation’ – a Sun Ultra 60. Only executable code was required, although some coders were provided as ‘C’ source code and compiled in-situ.
The use of three national test laboratories, each a specialist in speech testing with different testing methods, languages etc., ensures that the candidate coders for STANAG 4591 are exposed to a rigorous and varied testing process. Each of the national test laboratories provided the speech material required by their particular tests to the NC3A-NL. The NC3A-NL then processed this material through the candidate coders as described above. In Phase I, all coders were assessed in a limited number of noise environments e.g. quiet, modern office and two levels of non-stationary speech-shaped noise. The Phase I tests are summarised in Table 1 with the Signal to Noise Ratio for each of the four acoustic environments listed in Table 2.

The NC3A performed additional processing on the MOS test material before and after passing it through the candidate and reference voice coding algorithms [1].

<table>
<thead>
<tr>
<th>Test Characteristic Test lab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnostic Rhyme Test (DRT) [2] Intelligibility ARCON (US)</td>
</tr>
<tr>
<td>Consonant-Vowel-Consonant (CVC) [3][4] Intelligibility TNO (NL)</td>
</tr>
<tr>
<td>Intelligence Transmission (Inteltrans) [5] Intelligibility CELAR (FR)</td>
</tr>
<tr>
<td>Mean Opinion Score (MOS) [4] Quality ARCON (US)</td>
</tr>
<tr>
<td>Mean Opinion Score [4] Quality TNO (NL)</td>
</tr>
</tbody>
</table>

Table 1: Phase I test methods

<table>
<thead>
<tr>
<th>Condition Number</th>
<th>Noise condition</th>
<th>SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise 01</td>
<td>Quiet</td>
<td>$\infty$ dB</td>
</tr>
<tr>
<td>Noise 02</td>
<td>Speech noise</td>
<td>6 dB</td>
</tr>
<tr>
<td>Noise 03</td>
<td>Speech shaped noise</td>
<td>12 dB</td>
</tr>
<tr>
<td>Noise 04</td>
<td>Modern Office</td>
<td>20 dB</td>
</tr>
</tbody>
</table>

Table 2: Phase I noise conditions

In general, each speech file from the test laboratories was processed through nine voice coders – three candidate vocoders, each at two data rates, plus three reference coders. For calibration of the Mean Opinion Score (MOS) tests for both the US and NL test labs, the speech material was also passed through Modulated Noise Reference Unit (MNRU) software from the International Telecommunications Union (ITU) [6]. This adds known levels of noise to speech signals (in this case at 5 dB, 10 dB, 15 dB, 20 dB, 25 dB, 30 dB, 35 dB and 40 dB SNR).
PHASE I RESULTS

Phase one gives an indication of the levels of performance attainable with the state of the art voice coders being evaluated. Figures 1 and 2 show the results from one test laboratory, aggregated over all subjects. Figure 1 shows the intelligibility performance, figure 2 the speech quality metrics. The three candidate voice coders (A, B and C) are shown at 2400 bps and 1200 bps rates and in comparison to the three known references. It comparing candidates voice coders for STANAG 4591, it should be noted that only LPC-10e has a bit rate as low as 2.4 kbps; ACELP uses 4.8 kbps while CVSD uses 16 kbps.

![Figure 1: Phase 1 intelligibility results](image)
PHASE II TESTING

Phase II repeated all of the Phase I tests on the intelligibility and speech quality, but also performed these tests in a much wider range of acoustic noise conditions. These additional acoustic noise conditions are both harsher and more representative of the worse case NATO operational scenarios. These harsher acoustic noise environments include the Mobile Command Enclosure (MCE) field shelter, staff car, wheeled military vehicle (HMMWV and P4), helicopter (UH60 Black Hawk), tracked vehicle (M2A2 Bradley fighting vehicle and LeClerc tank) and supersonic aircraft (F-15 and Mirage 2000).

Tandem and random bit error testing for quality and intelligibility testing are two practical Phase II environments in which military voice coders operate. In the case of the ‘tandem’ condition, the speech passes through two complete voice coding algorithms. First, the 16Kbps CVSD algorithm followed by the candidate algorithms. The two voice coders in the tandem test are complete in that they each take speech in, write out digital bit streams (to files), and produce output speech. The random bit error channel test uses a 1% random bit error pattern applied to the digital bit stream files. The final condition is the performance of the candidates when presented with a whispered speech input signal. This test uses the Dutch (TNO) Speech Reception Threshold or SRT intelligibility test for evaluation.

Two additional types of tests are conducted in Phase II. These are the Speaker Recognisability and the Language Dependency Tests.

Speaker Recognizability Test – measures the ability of a listener to correctly determine the identity of the speaker using only the communications link under test. It is increasingly important
for the communicators to be able to identify the emotional state and identity of the field commanders and/or headquarters personnel over tactical links. This test seeks to measure this ability.

Language Dependency – This test again uses the Dutch (TNO) Speech Reception Threshold Test (SRT) to compare the performance of the candidate coders in several different languages. The NATO test includes English, French, Dutch, and German as the official languages under test. In the interest of scientific investigation, Polish and possibly Turkish may be included informally. The results of the SRT tests will be subjected to an Analysis of Variance Analysis (ANOVA) to identify whether there are candidates (or references) which perform poorly for a particular language or set of languages. As NATO is an organisation with 19 member nations and two official languages, with many non-native speakers attempting to communicate through the two official NATO languages to other non-native speakers, this is an important test.

**BLINDING PROCESS**

To mask the identities of all voice coders and guarantee impartiality during the analysis of individual voice coders, a blinding process is applied by the NC3A to all processed material before it is sent for analysis by the national test laboratories.

The output from each of the nine coders (three candidates, each at two bit rates, plus three reference coders) is randomly re-labelled as ‘coder n’ where \( n = 1 \) to 9. The test material is then sent to the national laboratories for appropriate evaluation. It should be noted here that with the application of the blinding process, the test laboratories have no information pertaining to the identity of the voice coding algorithms being evaluated.

In Phase I a single blind was applied by the NC3A host laboratory prior to evaluation of the data by the test laboratories. In Phase II, for added integrity, a double blind was carried out where the NC3A host laboratory performed the initial blinding, with a second blinding operation carried out by an impartial member of the NBVC AHWG. The double blind was carried out in isolation and provided a re-labelling of all ‘coder n’ to ‘vocoder m’. The result is that neither the NC3A personnel nor the impartial NBVC representative responsible for the second blinding operation are aware of the identities of the nine coders.

**SELECTION PROCESS**

The results of all tests, from all test laboratories are combined according to a pre-determined weighting matrix devised by AHWG NBVC [7]. Each of the tests is re-sampled to a common scale (based upon a statistical analysis of individual tests) with appropriate weights applied. This weighting affects the relative influence of the bit rate of the coders, the particular tests and acoustic noise conditions.

For both Phase I and Phase II, the relative weight of bit rate was 60% for 2400bps results and 40% for 1200bps results. For Phase I, within these bit rate divisions, weighting on different tests meant 55% of the score came from intelligibility results, and 45% from speech quality. These were further sub-divided with contributions from results in different noise conditions and from different test laboratories, etc.

Phase II has a more complex weighting matrix accounting for the additional tests and additional noise conditions. The initial calculation of the performance index used to combine the scores is done on completely double-blinded data identified only as vocoder 1 to \( m \). Once the preliminary calculation of the data has been disseminated to the members of the group the second level blinding will be removed at the October 2001 meeting to allow for the grouping of coder pairs and the identity of the references to be revealed. This procedure is necessary to allow the individual national labs to use the reference coder data to compute a series of checks and balances to ensure that no
mistakes were introduced into the testing process. Once the data for the 1.2Kbps and 2.4Kbps candidate pairs has been combined into final performance calculations the group will have an opportunity to thoroughly discuss the results. Only when the vocoder pair (1.2Kbps and 2.4Kbps candidate) with optimum performance in the conditions agreed by the NBVC AHWG is identified, will the final blind be removed to reveal the voice coder adopted as STANAG 4591, the future NATO narrow band voice coder.

PHASE III TESTING

Once all coders have been evaluated in Phase II, the candidate coder that offers the best overall performance will be known and the selection process for STANAG 4591 will be complete. Phase III is intended as a real time validation of the performance of the winning candidate under realistic operational conditions. Phase III will test the Communicability of the winning candidate for STANAG 4591. This test seeks to measure the efficiency of communication of STANAG 4591 when compared to a series of reference coders (as described above). This is accomplished by placing two live communicators in acoustic chambers connected together through a simulated communications channel. The communicators are asymmetrically subjected to realistic acoustic simulations including the following environments HMMWV, Blackhawk, E3A AWACS, F16 jet, MCE, etc. In effect, Phase III testing is attempting to make the most realistic simulation possible of a live communications link. The channel simulations will include HF, VHF, UHF and MILSATCOMs and will reflect the actual error conditions as seen by the voice coding algorithms in those particular scenarios. Bit perturbation patterns representing channel simulation data will be provided to the AHWG by the appropriate NATO sponsored working groups or designated authorities.

CONCLUSIONS

The introduction of a new NATO narrow band voice coder (STANAG 4591) will improve communications within NATO and enable end-to-end security and interoperability. Modern technology allows such voice coders to provide improved speech performance in a variety of harsh acoustic noise environments.

The test plan that was devised ensures that the voice coder to be selected as STANAG 4591 will provide these performance improvements. The rigorous and comprehensive nature of the testing plan ensures that these benefits will be available in the difficult military environments in which NATO forces must communicate.

The NATO standardisation of the state of the art STANAG 4591 voice coder ensures that the NATO and Partnership for Peace member nations will benefit from improved performance are maximised through combination with interoperability.

ACKNOWLEDGEMENTS

The authors would like to thank all the members of the NATO Ad-Hoc Working Group on Narrow Band Voice Coding AC322-SC/6-AHWG/3, past and present, for their contributions to this work. The authors are also grateful to those non-AHWG members within the national test laboratories who have supported this work.

REFERENCES


