Evaluation of Low Rate Speech Coders for HF

Thomas E. Tremain, David P. Kemp, John S. Collura, Mary A. Kohler

U.S. Department of Defense

Abstract

In 1988, the NATO Tri-Service Group for Communications and Electronics Equipment (TSGCEE) Sub Group 11 (SG/11), established Working Group 2 (WG/2, Narrowband Speech), to develop a voice processor standard for a secure voice system for operation in the High Frequency (HF) portion of the electromagnetic spectrum. The voice processors evaluated had a bit rate of 600, 800, 900, and 1200 bps. Forward error correction was added to bring the total bit rate up to 2400 bps [1]. This paper compares the performance of the three candidate U.S. low rate speech coders, to the standard 2400 bps Linear Predictive Coder, under a number of different test conditions.

I Proposed Test Procedures

After several meetings, WG/2 established a process for selection of the speech algorithm to become the NATO Standard. Each country wishing to submit a voice processor operating at any of the above bit rates notified WG/2 at its September 1991 meeting. Performance testing was accomplished as follows:

- Each country wishing to submit candidates identified (to WG/2) the voice processors it will submit.

- Each country wishing to submitted a test plan to WG/2 for performing the tests. WG/2 specified modifications to the test plans as necessary.

-Each country submitting a test supplied all of the countries submitting candidates with the approved test plan and the associated database to be used to perform the tests on the candidates.

- Test results were returned to the country of origin and evaluated in accordance with its approved test plan.

The performance evaluations along with complexity and delay of the voice coders were presented at the September 1992 meeting of WG/2. Real-time hardware versions of the candidate algorithms were also demonstrated at this meeting. WG/2 will make its final selection based on the performance evaluations, complexity, delay and demonstrations of the real-time hardware.

II Candidate Coders

Belgium submitted two coders running at 900 bps. The low rate coders were based on the Multiband Excited Speech Coder.

France submitted three coders, two running at 800 bps and one at 1200 bps. The low rate coders were based on the Linear Predictive Speech Coder [2].

The United States submitted three candidates, one each at 600, 800 and 1200 bps. The low rate coders were based on the Linear Predictive Speech Coder [3], [4].

All candidates were compared to the present NATO standard which is Linear Predictive Coding at 2400 bps.

III Tests Run on Speech Coders

U.S. Tests

The U.S. source material consisted of digital recording of the Diagnostic Rhyme Test (DRT)[5] and the Diagnostic Acceptability Measure (DAM)[6] test, for each of the four microphone and acoustic background combinations and two error conditions as listed in Table 1. The DRT is a two choice intelligibility test in which each item consists of two rhyming words, selected so that the initial consonants differ by a single phonetic attribute, for example vault vs. fault. The listeners task is simply to judge which of the two words have been spoken. The test has 192 words which are scored and 40 unscored filter or test words. The final score is the number of correct words minus the number of incorrect words divided by the total number of words. The single speaker DRT lasts approximately 7 minutes.

The DAM speech data base per speaker consists of 12 sentences spoken at a rate of one sentence every 4 seconds. A personal computer is used to record the responses of a listener to nine signal distortions, eight background noise distortions and three overall effects, intelligibility, pleasantness and acceptability. Each distortion is rated on a scale of (0) - not detected to (9) - overwhelming. The source material contained 3 male and 3 female speakers.

Table 1: U.S. Test Conditions

Test No	Microphone	Acoustic Background	Error Rate	Weight
1	Dynamic	quiet	0% ber	25%
2	Vinson	quiet	0% ber	15%
3	H250 (nc)	jeep	0% ber	15%
4	EV985 (nc)	tank	0% ber	15%
5	dynamic	quiet	1% ber	15%
6	dynamic	quiet	3% ber	15%

Netherlands Test

The Netherlands source material consisted of Consonant Vowel Consonant (CVC) digital recordings for the conditions listed in Table 2. For the Dutch language this results in 17 initial Consonants (C_I), 15 Vowels (V) and 11 final Consonants (C_F). Each word list consists of 51 CVC combinations, resulting in both nonsense as well as meaningful words. The CVC words are embedded in five different carrier phrases, with each speaker reading 50 different word lists of embedded CVC words. There are a total of 400 different lists derived from permuting carrier phrases with nonsense words. Each phrase was presented separately in a sequence of one phrase every 3 seconds. Hence, a list of 51 words and phrases lasts a total of 153 seconds.

Table 2: Netherlands Test Conditions

Test No.	Acoustic Noise
1	Quiet
2	Noise Condition 6 -12db S/N
3	Noise Condition 6 -6db S/N
4	Noise Condition 14 -12db S/N
5	Noise Condition 14 -6db S/N

The source material contained 4 male and 4 female speakers. The two noise conditions are part of the NATO RSG-10 Noise Data Base. Noise condition 6 is speech noise with an average speech spectrum. Noise condition 14 is the noise present in the operations room of a destroyer.

French Test

The French digital source material consists of a corpus of 216 words all uttered by 2 males and 1 female. The conditions are listed in Table 3. The French test is the Diagnostic Rhyme Test as defined in [5] and adapted to the French language by M. Rosey and Cartier. The final score on the French DRT is the number of correct words divided by the total number of words. The background noise is taken from the NATO RSG-10 Noise CD-ROM. The noise selected is the operation room noise.

Table 3: French Test Conditions

Test No.	Microphone	Acoustic Background	Error Rate	Weight
1	dynamic	quiet	0% ber	25%
2	dynamic	tactical noise	0% ber	25%
3	military	quiet	0% ber	25%
4	dynamic	quiet	3% ber	25%

Canadian Test

The Canadian tests are the Mean Opinion Score (MOS) test and the Degraded Mean Opinion Score (DMOS) test. The MOS test is mainly for quiet conditions while the DMOS test is mainly for the acoustic noise conditions. Listeners are asked to judge the low rate coders relative to the 2400 bps LPC-10e coder. The MOS scale is [1-bad, 2-poor, 3-fair, 4-good and 5-excellent]. The DMOS scale is [1-very much poorer quality, 2-much poorer quality, 3-moderately poorer quality, 4-slightly poorer quality, and 5-the same or better quality]. The Canadian digital source material consists of 350 MOS sentence files. Table 4 shows the test conditions. The source material is based on four speakers. Each file contains two sentences.

Table 4: Canadian Test Conditions

Test No.	Acoustic Background
1	Quiet, (270 files)
2	Helicopter S/N = 1dB, (20 files)
3	Helicopter S/N = 10dB, (20 files)
4	Armoured Personnel Carrier (APC) - S/N=1dB, (20 files)
5	APC - S/N = 10db, (20 files)

IV Tests Results

Table 5 compares the ratings, scores and standard deviations for each of the tests. Note that the French DRT test on all the coders was not completed before the September meeting.

Table 5: Test Rating Comparison

Category			Scores			
Speech Rating	Degradation	Listening Effort Required	CVC	DRT	DAM	MOS/ DMOS
Excellent	None	None	>97	>96	>75	5
Good	Slightly poorer	Little Attention	85-97	87-96	60-75	4
Fair	Moderately poorer	Moderate	65-85	79-87	45-60	3
Poor	Much poorer	Considerable	33-65	70-79	30-45	2
Bad	Very much poorer	Very hard to understand	<33	<70	<30	1
Test	Standard	Deviation	2.5	1.0	1.5	0.15

Quiet Input Conditions: Table 6 shows the quiet Intelligibility and Quality test results for the different U.S. speech coders compared against the reference Linear Predictive Coder at 2400bps. The intelligibility test results (DRT) and (CVC) and the quality test results (DAM), (MOS) and (DMOS) show that a coder at 800bps provides performance almost as good as a 2400 bps coder. Based upon channel simulations where the 600, 800 and 1200bps voice coders are error corrected to 2400bps, the preferred data rate is 600bps or 800bps because 1200bps with a half rate code does not provide acceptable performance over a degraded HF channel. The performance in random bit errors between the 600bps and 800bps coders is small as shown in tables 11 and 12. However there appears to be a large difference in performance between the 600bps and 800bps coders based upon intelligibility and quality tests measured in various acoustic noises and the quiet background condition as shown in tables 6, 7, and 8.

Table 6: Quiet test conditions

Quiet		LPC-10e	600	800	1200
US DRT	Male	92.9	89.5	90.9	92.4
	Female	86.8	84.5	86.3	85.2
	Combined	89.9	87.0	88.6	88.8
US DAM	Male	57.0	52.2	55.2	54.7
	Female	54.2	49.6	52.5	53.1
	Combined	55.6	51.0	53.9	54.0
CVC	Male	66.9	48.4	64.0	66.8
	Female	65.5	47.9	56.4	54.8
	Combined	66.2	48.2	60.2	60.8
MOS/ DMOS	Male	4.0/	3.2/4.3	3.9/4.9	3.9/4.8
	Female	3.4/	3.1/4.4	3.8/4.8	3.8/4.9
	Combined	3.6/	3.15/4.4	3.85/4.85	3.85/4.85

Intelligibility in Acoustic Noise: Table 7 shows the U.S. speech coder intelligibility test results based on the U.S. DRT test and the Netherlands CVC test. For the US DRT test jeep and tank noises were added acoustically using a first order gradient noise cancelling microphone. The S/N was improved by 12-15 dB by using the noise cancelling microphone. For the Netherlands CVC test the average speech spectrum noise and destroyer operations room noise were added electrically at -6dB and -12dB relative to speech level. Since both noise environments have most of the noise in the lower frequencies the S/N would be improved 12-15 dB by using a noise cancelling microphone. The U.S. test was run at Dynastat in Austin Texas and the Netherlands test was run at TNO in Soesterberg the Netherlands.

Table 7: Intelligibility Test Conditions for Acoustic Noise

Acoustic	Noise	LPC-10e	600	800	1200
DRT Tank	Male	80.9	77.8	77.6	80.5
	Female	76.5	72.8	73.3	74.6
	Combined	78.7	75.3	75.5	77.6
DRT Jeep	Male	86.7	81.2	84.5	86.1
	Female	79.9	76.1	77.8	77.6
	Combined	83.3	78.6	81.2	81.9
CVC Speech Spect -6dB	Male	43.0	19.9	28.7	36.0
	Female	24.8	17.8	20.7	21.0
	Combined	33.8	18.9	24.7	28.5
CVC Operations Rm -6dB	Male	27.9	17.6	24.4	27.6
	Female	15.4	15.3	13.2	14.8
	Combined	26.7	16.4	18.8	21.2
CVC Operations Rm -12dB	Male	49.3	33.0	39.7	50.2
	Female	34.2	21.6	31.9	33.3
	Combined	41.7	27.3	35.9	41.8
CVC Speech Spect - 12dB	Male	53.9	33.5	48.9	53.7
	Female	38.0	27.9	30.8	39.5
	Combined	46.0	30.7	39.8	46.5

Quality in Acoustic Noise: Table 8 shows the U.S. speech coder quality test results based on the U.S. DAM test and the Canadian MOS and DMOS test. The U.S. test was run at Dynastat and the Canadian test was run at BNR in Ottawa, Canada. For the U.S. DAM test the jeep and tank noise were acoustically added using a first order gradient microphone. For the MOS/DMOS test the speech was preemphasized and the Helicopter and Armoured Personnel Carrier noise was added electrically at -1 dB and -10 dB relative to the speech level. The Canadian MOS/DMOS test compared the quality of the low rate coders against the LPC-10 at 2400 bps.

Table 8: Quality test for Acoustic Noise

Acoustic Noise		LPC-10e	600	800	1200
DAM Jeep	Male	41.5	38.4	40.5	41.1
	Female	41.4	38.0	39.1	41.1
	Combined	41.4	38.2	39.7	41.0
DAM Tank	Male	42.2	38.8	40.8	40.9
	Female	37.8	37.0	37.5	39.0
	Combined	40.0	37.9	39.1	40.0
MOS/DMOS APC -1dB	Male		/3.8	/3.9	/4.2
	Female		/3.8	/4.1	/4.2
	Combined	1.5/	1.3/3.8	1.4/4.0	1.5/4.2
MOS/DMOS APC -10dB	Male		/3.8	/4.1	/4.3
	Female		/4.2	/4.3	/4.5
	Combined	1.95/	1.6/4.0	1.9/4.2	2.0/4.4
MOS/DMOS Helicopter -1 dB	Male		/3.85	/4.0	/4.1
	Female		/3.75	/4.0	
	Combined	1.4/	1.2/3.8	1.3/4.0	1.4/4.2
MOS/DMOS Helicopter -10dB	Male		/4.0	/4.1	/4.1
	Female		/4.2	/4.8	/4.5
	Combined	1.9/	1.7/4.1	1.8/4.3	2.1/4.3

Comparison of flat with tactical microphone: Table 9 compares the Intelligibility (DRT) for a flat microphone and a tactical microphone which preemphasizes the input with a first order digital filter approximated by $1-0.9z^{-1}$. Table 10 compares the quality (DAM) for a flat microphone with a tactical microphone. There is a loss in intelligibility and quality for the tactical microphone.

Table 9: Microphone Intelligibility Comparison

US DRT		LPC-10e	600	800	1200
Flat Microphone	Male	92.9	89.5	90.9	92.4
	Female	86.8	84.5	86.3	85.2
	Combined	89.9	87.0	88.6	88.8
Tactical Microphone	Male	90.8	87.7	89.0	90.8
	Female	86.7	81.4	83.9	86.9
	Combined	88.7	84.5	86.4	88.9

Table 10: Microphone Quality Comparison

US DAM		LPC-10e	600	800	1200
Flat Microphone	Male	57.0	52.2	55.2	54.7
	Female	54.2	49.6	52.5	53.1
	Combined	55.6	51.0	53.9	54.0
Tactical Microphone	Male	47.4	44.6	46.3	46.0
	Female	47.1	47.4	49.5	48.9
	Combined	47.3	46.0	47.9	47.4

Testing in random bit errors: For this particular test there was no error protection for the 1200 bps voiced coder. The 600 bps and 800 bps voice coders were error corrected to 1200 bps. The 2400 bps standard LPC-10 coder has some error correction during unvoiced speech and uses smoothing on the voiced speech parameters when errors are detected during the unvoiced portions of speech.

Table 11: Intelligibility Comparison in Errors

US DRT		LPC-10e	600	800	1200
1% BER	Male	90.9	89.8	88.1	87.1
	Female	85.4	85.5	86.6	83.8
	Combined	88.1	87.6	87.4	85.4
3% BER	Male	86.0	87.8	87.4	80.4
	Female	81.3	84.1	83.2	79.4
	Combined	83.7	85.9	85.2	79.9

Table 11 gives the DRT results for the coders with 1 and 3 percent random bit errors. Comparing the results in table 11 with the quite condition results of table 6 shows the coders that provided the best performance are the 600 and 800 bps coders which have error protection.

Table 12: Quality Comparison in Errors

US DAM		LPC-10e	600	800	1200
1% BER	Male	47.4	50.3	52.4	48.2
	Female	49.1	47.4	52.2	46.4
	Combined	48.3	48.9	52.3	47.3
3% BER	Male	42.2	52.5	54.1	39.5
	Female	41.5	48.1	52.4	36.4
	Combined	41.9	50.3	53.2	37.9

Table 12 gives the DAM results for the coders with 1 and 3 percent random bit errors. Comparing these results with those for the quiet condition found in table 6 shows that the 800 bps coder has almost no loss in quality between the quiet and the 3 percent case. The 1200 and 2400 bps coders have a DAM loss of about 15 points between the quiet and 3 percent case.

IV Conclusions

All of the test results reported in this paper were reviewed at the NATO Working Group 2 meeting in September 1992. Under the majority of the test conditions, the 800 bps speech coder is performing close to the present standard LPC-10 at 2400 bps. The one area where there is a loss of performance in the 800 bps speech coder relative to the LPC-10 is in the severe CVC intelligibility test conditions. All of the tests will be completed and a coder at 800 bps will be selected at the WG/2 meeting in March 1993.

References

[1]A.C. Duke, D.J. Rahikka, "Error Correction of Low Rate Speech in Jammed Frequency Hop H.F." Proc. Tactical Communications Conference TCC 92.

[2]B. Mouy P. de la Nove, "Voice Transmission at a very Low Bit Rate on a Noisy Channel," Proc. IEEE ICASSP 1992 Vol. 2, pp. 149-152.

[3] D.P. Kemp, J.S. Collura, and T.E. Tremain "Multi-Frame Coding of LPC Parameters at 600-800 bps," Proc. IEEE ICASSP 1991 pp. 609-612.

[4] D.P. Kemp, "LPC Parameter Quantization at 600, 800, 1200 bps,"Proc. Tactical Communications Conference, TCC 92

[5]W.D. Voiers, "Diagnostic Evaluation of Speech Intelligibility." In Speech Intelligibility and Speaker Recognition, Vol. 2 pp. 374-394.

[6]W.D. Voiers, "Diagnostic Acceptability measure for Speech Communication Systems," Proc. IEEE ICASSP, 1977 pp. 204-207.