

# VOCODER INTELLIGIBILITY AND QUALITY TEST METHODS

*John D. Tardelli and Elizabeth Woodard Kreamer*

ARCON Corp., 260 Bear Hill Road Waltham, Massachusetts 02154  
email jdt@arcon.com

## ABSTRACT

The U.S. Department of Defense's Digital Voice Processing Consortium (DDVPC) is in the process of choosing a new 2400bps vocoder standard. As part of the selection process, a study was conducted to determine the test measures to be used for the evaluation of the intelligibility and quality of competing vocoder algorithms. The DDVPC requirements for a vocoder differ substantially from commercial interests. Severe ambient acoustic noise backgrounds coupled with hostile transmission channel conditions must be considered. The commercially popular Mean Opinion Score (MOS) test method was evaluated as a replacement or companion test of quality to the Diagnostic Acceptability Measure (DAM), traditionally used by the DDVPC. The study showed that a properly structured MOS test can achieve equal resolution, reliability, and validity to that of the DAM at equivalent costs. Any MOS test series must be structured to minimize contextual effects. Certain severe conditions must use the Degraded Mean Opinion Score (DMOS) test method to achieve usable resolution.

## 1. INTRODUCTION<sup>1</sup>

The U.S. D.o.D. Digital Voice Processing Consortium (DDVPC) is in the process of choosing a new 2400bps vocoder standard [1]. As part of the selection process, a study was conducted to determine the test measures to be used for the evaluation of the intelligibility and quality of competing vocoder algorithms. The DDVPC has extensive experience in the selection of voice coders and has traditionally used the Diagnostic Rhyme Test (DRT) as a measure of intelligibility and the Diagnostic Acceptability Measure (DAM) as a measure of quality. During the recent increased commercial interest in narrowband vocoders, other test measures have become popular. The DDVPC requirements for a vocoder differ substantially from commercial interests. Severe ambient acoustic noise backgrounds coupled with hostile transmission channel conditions must be considered. A study was undertaken to evaluate the commercially popular Mean Opinion Score (MOS) test method as a replacement or

companion test of quality to the DAM. Controlling the extensive cost of running a vocoder selection test for the multiple conditions of interest to the DDVPC was the objective of this study. The study showed that a properly structured MOS test can achieve equal resolution, reliability, and validity to that of the DAM at equivalent costs. Any MOS test series must be structured to minimize contextual effects. Certain severe conditions must use the Degraded Mean Opinion Score (DMOS) test method to achieve usable resolution. The study was baseline against an Experts Opinion A/B Comparison study. These results indicated a problem with Expert Bias for those algorithm developers who limited their work to a single type of vocoder algorithm. The DDVPC chose the MOS/DMOS as the test method to be used for the 1995/1996 2400bps vocoder selection test.

## 2. INTELLIGIBILITY MEASURE

Intelligibility testing is used to determine if the speech being transmitted is able to be interpreted and, therefore, understood. The most common approach for intelligibility determination is the use of the DRT. This test assumes that if a coded word, or actually, list of words is perceived correctly, the system is intelligible. The DRT uses rhyming word pairs that vary only in the first consonant. The listener is presented visually with the word pair and then aurally with a member of that specific pair. The task is to select from the pair the one that was presented aurally. The scoring procedures are clear in that it is known which word was spoken and which was selected as heard by the listener. By calculating the number of words in the list that were perceived correctly it is possible to calculate a score for the level of intelligibility of the system.

The DDVPC and USAF RL/ERC-1 Office have an extensive library of prerecorded DRT word lists for a large variety of talkers in numerous acoustic noise environments. This library of input test material incorporates resident microphone characteristics and talker responses to the acoustic noise by recording the test material in simulated noise fields. The entire library has also been level equalized to eliminate vocoder overflow and presentation level problems when testing. The DDVPC Test Plan [2] calls for the evaluation of intelligibility in a large number of acoustic noise environments. The DDVPC's past success with the DRT in

---

<sup>1</sup>This work is supported by the US Air Force RL/ERC-1 Office, Contract #F19628-C-95-0190.

selection tests and the extent of the source material library, led to the DRT being chosen as the intelligibility measure for the new 2400bps selection process.

### 3. QUALITY MEASURE

Quality testing often is used to augment or take the place of intelligibility testing. It provides a picture of the personal opinions of the listeners regarding the signal as transmitted by the communication systems or processed by the algorithms being tested. As normally conducted, subjective studies of speech communication systems require evaluators to listen to processed speech for a defined period of time. The evaluators then may be asked to make ratings regarding system quality as with the Mean Opinion Score (MOS) or regarding the impact of the system on various communication quality attributes as with the Diagnostic Acceptability Measure (DAM), two of the most commonly used quality measures.

The commercial world has been relying on MOS testing rather than the DAM, while historically, the DAM-IIA was the primary vehicle for subjective evaluations of coder diagnostics and for coder selection within the DoD. Recently there has been a significant overlap of DoD and commercial interests. This overlap, coupled with the knowledge that a new version of the DAM (DAM-IIC) had recently been implemented by Dynastat Inc., the developer of the DAM, led to a decision by the DDVPC Test and Evaluation (T&E) Committee to investigate the reliability of MOS testing and the ability of both the MOS and DAM tests to adequately indicate differences in coders.

#### 3.1 DAM/MOS Comparative Study

The study utilized ten vocoder algorithms under three conditions, three reference conditions and seven calibration systems. Three male and three female talkers were used with all systems. The three conditions were: 1) Quiet environment with Dynamic mic and clear channel, 2) Jeep environment with Electro Voice H250 mic and clear channel, 3) Quiet environment with Dynamic mic and 1% random bit error channel. The three reference conditions were: 1) unprocessed Quiet, 2) unprocessed Jeep, 3) a Plain Old Telephone (POTS) simulation of the Quiet input. The MNRU degradation at Q-values of 5,10,15,20,25,30,35 was used for calibration. The vocoders ranged in data rate from 2400bps to 32Kbps. They were: ADPCM, CVSD32, CVSD16, VSELP, CELP, STC-4.8, MBE-4.8, LPC-10E, STC-2.4, MBE-2.4. DAM tests were conducted at Dynastat in Austin, TX. MOS testing was conducted at two laboratories. In addition to the full test, a limited "worst-case" context experiment was also held at the MOS laboratories. The first MOS experiment consisted of 240 test items separated into six blocks of 40 items each. The second experiment consisted of 72 test items separated into four blocks of 18 items each. Forty listeners were used for each experiment at each laboratory. The MOS laboratories were COMSAT in Clarksburg MD and MPR/Simon Fraser University in Vancouver, BC.

Since the DAM historically has been the test method used to select digital voice processor algorithms for the DoD, much more was known about its reliability and validity. Therefore, the determination was made to place the heaviest burden of proof on the MOS. Specifically, if the MOS did not prove to be reliable across test laboratories, the DAM would be chosen. The analyses were conducted to determine the efficacy of each test according to the criteria of cost effectiveness, reliability, validity, and high resolution.

Cost effectiveness was determined a priori in that neither test, as used in this study, was deemed to be significantly more expensive than the other. Reliability for the MOS was determined in part by comparing the results from the two different laboratories. This allowed the investigation of the impact of inherent differences (e.g. subject pools, equipment, etc.) on MOS test reliability. Although critical, this is only one aspect of reliability. Context effects are a potential form of MOS unreliability and were addressed as such. This was evaluated both within and across laboratories and was found to be a concern. Results indicated that the quality of the systems tested does have an impact on the scores produced by MOS test. Further, based on these results it appears that male and female listeners may differ on their ratings for communication systems. To negate this effect, it seems imperative that MOS listener crews be comprised of an equal number of males and females and that MOS test design must minimize context effect.

The repeats of calibration and reference conditions for the DAM-IIC indicated that the DAM-IIC demonstrates test-retest reliability. Although there are intra-session score corrections built into the DAM, the wide range of the scores intimated that there might be a problem in making direct comparisons between systems. Finally, there was a suggestion that the DAM-IIC was having difficulty differentiating between systems. However, these observations are qualified by the fact that the analyses were limited due to the summary form of the data provided for the DAM-IIC and the number of systems it was possible to test.

Both tests produced scores that met logical validity tests. Test results averaged across all Talkers are presented for both MOS laboratories and the DAM in Tables 1 and 2. The solid bars to the right of the scores indicate 95% Confidence Interval equivalent groups as calculated using the Newman-Keuls method of equivalent differences. The coders used in this study were generic implementations and had some problems. These scores should not be used as a basis of coder algorithm selection. Complete results can be found in [3].

Table 1. presents MOS and DAM Newman-Keuls comparisons based on results from Experiment 1 for combined Talkers in the Quiet. Observation of the absolute rankings for the two sets of MOS results shows that in only one instance do they differ from each other. CVSD32 and STC24A are reversed. Using the error bars, it is possible to see that the ranks are equivalent. Thus, results from the two

MOS laboratories demonstrate similar degree of resolution and rankings are the same or equivalent.

Table 1. MOS/DAM Newman-Keuls Comparisons  
All Talkers, Quiet, Experiment #1

MOS (COMSAT)	MOS (MPR)	DAM-IIC
NULL 4.38	NULL 4.25	NULL 79.8
ADPCM 3.90	ADPCM 3.59	ADPCM 69.0
VSELP 3.52	VSELP 3.16	VSELP 66.5
STC48 3.08	STC48 2.83	CELP 59.9
MBE48 2.99	MBE48 2.76	MBE48 57.8
CELP 2.98	CELP 2.70	STC48 56.2
	STC24B 2.57	
STC24B 2.72	STC24A 2.48	STC24B 53.2
CVSD32 2.59	CVSD32 2.45	STC24A 51.9
STC24A 2.59		MBE24 51.1
	MBE24 2.15	CVSD32 50.0
MBE24 2.34		LPC10E 49.4
	LPC10E 1.98	
LPC10E 2.13	CVSD16 1.90	CVSD16 43.2
CVSD16 1.93		

A comparison of the ranks from the MOS tests in the Quiet conducted by MPR with those from the DAM-IIC shows that there are rank differences involving four coders (STC48, CELP, CVSD32, and MBE24). Of these, it is notable that the MOS laboratories agree in ranking STC48 and MBE48 above CELP, while the DAM-IIC differs by ranking CELP above them both. Using the error bars and looking at ranks for STC48 and CELP, the MOS ranks are equivalent to those for the DAM-IIC while the ranks from the DAM IIC are not equivalent to those for the MOS. For CVSD32 and MBE24, the MOS ranks are not equivalent to those for the DAM-IIC while the ranks from the DAM IIC are equivalent to those for the MOS.

Comparing the results from the MOS tests in the Quiet conducted by COMSAT with those from the DAM-IIC shows that there are rank differences involving five coders (STC48, CELP, CVSD32, STC24A, and MBE24). The error bars illustrate that the two tests are equivalent for STC24A. As with the results from MPR, STC48 and CELP MOS ranks are equivalent to those for the DAM-IIC and the ranks from the DAM IIC are not equivalent to those for the MOS. With CVSD32 and MBE24, the MOS ranks are not equivalent to those for the DAM-IIC while the ranks from the DAM IIC are equivalent to those for the MOS.

Table 2. presents MOS and DAM Newman-Keuls Comparisons based on results from Experiment 1 for combined Talkers in the Jeep acoustic background. Observation of the absolute rankings for the two sets of MOS results shows that they differ from each other in only one instance. In this environment, LPC-10E and MBE24 are reversed. Using the error bars, it is possible to see that the ranks are equivalent. Thus, results from the two MOS laboratories demonstrate rankings are the same or equivalent.

Table 2. MOS/DAM Newman-Keuls Comparisons  
All Talkers, Jeep, Experiment #1

MOS (COMSAT)	MOS (MPR)	DAM-IIC
NULL 2.97	NULL 2.76	NULL 56.5
ADPCM 2.85	ADPCM 2.65	ADPCM 55.2
VSELP 2.52	VSELP 2.27	VSELP 50.6
STC48 2.18	STC48 1.99	CELP 46.3
CELP 2.10	CELP 1.93	
MBE48 1.98	MBE48 1.90	MBE48 44.3
STC24B 1.92	STC24B 1.79	STC48 43.3
STC24A 1.80		STC24B 42.4
CVSD32 1.80	STC24A 1.58	
	CVSD32 1.57	STC24A 39.6
LPC10E 1.57		MBE24 39.4
MBE24 1.55	MBE24 1.44	LPC10E 37.9
CVSD16 1.50	LPC10E 1.36	CVSD32 36.4
	CVSD16 1.32	CVSD16 33.5

Using Table 2., a comparison of the ranks from the MOS tests in the Jeep acoustic background conducted by MPR with those from the DAM-IIC shows that there are rank differences involving six coders (STC48, CELP, MBE48, CVSD32, MBE24, and LPC-10E). The error bars illustrate that the two tests are equivalent for the MBE48 and LPC10E. STC48 and CELP MOS ranks are equivalent to those for the DAM-IIC while the ranks from the DAM IIC are not equivalent to those for the MOS. The MBE24 MOS rank is not equivalent to the DAM-IIC while the rank from the DAM IIC is equivalent to that for the MOS. CVSD32 rank is not equivalent to the DAM-IIC nor is the rank from the DAM IIC equivalent to that for the MOS. The same results are seen when the COMSAT data is compared to the DAM-IIC.

#### 4. EXPERTS TEST

It was recognized from the onset of this study that a possible outcome of the analyses was the results of the two MOS laboratories would be comparable, but would differ from those for the DAM. The question then would be which laboratory's set of rankings was "right." Though the systems had been chosen specifically to provide predictable outcomes, these outcomes were not absolute. It was decided to try to reduce the uncertainty that would occur with such an outcome by conducting a parallel effort aimed at determining a valid or "true" ranking for these systems.

The ultimate goal of either the DAM or the MOS is to reliably, and with a large degree of validity, predict the actual coder user preference for a system. While it is possible to demonstrate the reliability of a test with relative ease, demonstrating validity is not as straight forward. One standard approach for validating a test is to use an instrument that has been shown to be valid through time and/or experience. An evaluation then is conducted where both tests are administered and the results of the second instrument are statistically compared to the "true" results of the first. The obstacle for using that approach for this evaluation was that there were questions regarding the validity (appropriate

ranking of systems for coder selection) for both the DAM and the MOS and no other instrument exists to provide a test standard.

A study was designed which would allow the comparison of "true" rankings to those produced by each of the tests. Because it was felt that there would be differences in the opinions even experts have regarding the ranking of the systems and conditions chosen for this evaluation, this approach involved first obtaining an independent listing of the "true" rankings. This was accomplished through the use of an A/B comparison test where participants were forced to choose the preferred system. The forced choice between an A/B presentation meant that they were not allowed to indicate that they could not differentiate. Thus, there were no ties. They were given the option of self-pacing the test or listening to a pair more than once prior to making a decision.

#### **4.1 Expert A/B Opinion Test Procedures**

An A/B test approach can be very time consuming for the participants. Including all possible pairings of systems, conditions, and speakers would have required volunteer participants to expend considerable effort and time on the project. It was felt that this would make it unlikely that any of the targeted individuals would have been willing to participate. Therefore, only a single male speaker (JE), and a single condition (Quiet) was included in the study. To further reduce the amount of pairs only the processed and unprocessed systems of interest were used. MNRUs and unprocessed reference conditions were not included, nor was order of the presentation of the systems (i.e. A/B versus B/A) considered. A subset of systems from Experiment 1 of the MOS comparisons were selected for this study.

#### **4.2 Expert Listener Factor**

Although tangential to the current project, the data from the experts test were examined to determine if there was an effect due to the occupation of the expert. Specifically the question was asked if there was a bias for a coder developer for the system with which they were the most familiar. The answer was an unequivocal "yes."

Prior to initiating this effort, it was hypothesized that Unprocessed Speech would rank first and that it would be possible to differentiate between Unprocessed and all Processed Speech. Further, the expected rankings were that unprocessed quiet, ADPCM, and VSELP would be first, second and third respectively with the rest of the systems somewhere under those. Only twelve individuals produced the anticipated ranking of unprocessed quiet, ADPCM, and VSELP. Of the 41 participants, 28 had some other coder in the third or higher position. Five individuals rated a coder

equal to or better than unprocessed speech. Most importantly, of the ten coder developers in the study, all ranked their coder third or better. Three of this ten ranked their coder better than unprocessed speech. This is not to suggest that any of the coder developers were deliberately trying to rate their system high. Conversations with various of the raters after the study was completed indicated that they felt that they were being harsher with their own system.

The following analogy may help to explain this effect. Individuals moving to a new area become acclimated to the sound of a regional accent or dialect and eventually accept it as normal. The coder developers may become so familiar with the sound of the peculiarities of their own system they accept it as normal. Further, just as someone traveling to another region is acutely sensitive to minor accent differences from their own, the coder developers may be overly sensitive to differences in speech that is not from the coder they are the most familiar with; their own. The results serve to illustrate the potential problems inherent in having raters that are overly familiar with any one system.

## **5. CONCLUSIONS**

The MOS and the DAM-IIC appear to have had the same resolving capability between coders. The main difference was in the actual rank order of the coders. Several differences in equivalent rank order were found for the MOS and DAM-IIC. The most consistent and worrisome was the reversal in the ranking of the STC48 and CELP coders. For these coders, the Experts A/B test agreed with the MOS ranking where the STC48 coder ranked above the CELP coder. This result agrees with the APCO-25 test series where similar STC and CELP coders were compared. Because of these agreements in rank order, it was recommended that the MOS test method be adopted by the DDVPC as the preferred method for coder selection based on speech quality.

## **6. REFERENCES**

- [1] Tremain, Kohler, Champion, "Philosophy and Goals of the D.o.D. 2400bps Vocoder Selection Process", Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, May 1996.
- [2] Bielefeld, Supplee, "Developing a Test Program for the D.o.D. 2400bps Vocoder Selection Process", Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, May 1996.
- [3] Tardelli, Kremer, La Follette & Gatewood, "A Systematic Investigation of the Mean Opinion Score (MOS) and the Diagnostic Acceptability Measure (DAM) for Use in the Selection of Digital Speech Compression Algorithms", ARCON Corp. Sept. 1994.