A COMPARISON OF THE NEW 2400 BPS MELP FEDERAL STANDARD WITH OTHER STANDARD CODERS

M. A. Kohler

U.S. Department of Defense, 9800 Savage Road STE 6516, Ft. Meade, MD 20755-6516
makohle@alpha.ncsc.mil

ABSTRACT

In 1996, the U.S. Department of Defense Digital Voice Processing Consortium (DDVPC) selected Texas Instrument's mixed excitation linear prediction (MELP) algorithm as the recommended new federal standard for 2400 bps voice communications. The algorithm selection process involved quality, intelligibility, communicability, and recognizability testing in many acoustic noise, error, and tandem conditions. Algorithm complexity was also measured. This paper compares the performance scores, diagnostic information, and complexity of MELP to the 4800 bps federal standard (FS1016) code excited linear prediction (CELP) algorithm, the 16 kbps continuously variable slope delta modulation (CVSD) algorithm, and the venerable federal standard (FIPS Pub. 137) 2400 bps linear predictive coding (LPC-10) algorithm.

INTRODUCTION

At the request of civilian and military 2400 bps equipment users, the U.S. Department of Defense Digital Voice Processing Consortium (DDVPC) conducted a three year evaluation to recommend a replacement for 2400 bps LPC-10. The Consortium's goal for the new algorithm was that it meet or exceed the performance of CELP. This process culminated in March 1996 when the DDVPC chose the mixed excitation linear prediction (MELP) algorithm [1] as this replacement. Performance and complexity were the criteria.

Four performance tests were conducted: quality, intelligibility, recognizability, and communicability. Twenty-three noise conditions were tested across these four tests. All performance tests used equal numbers of male and female talkers. Evaluated processor and memory usage comprised the complexity measurement [2].

Three reference coders were similarly tested:
- 4800 bps code excited linear prediction (CELP) (FS1016) [3]
- 16 kbps continuously variable slope delta modulation (CVSD)
- 2400 bps linear predictive coding (LPC-10e) [4]

This paper provides graphical combined talker score comparisons. Error bars are also plotted representing the standard error. The symbols representing individual coders remain consistent throughout the paper and are listed on each plot. Scores for each coder in all testing conditions are shown when feasible. Abscissa values are the MELP scores for each condition.

QUALITY

Quality testing used the Mean Opinion Score (MOS) for benign noise conditions, and the degradation Mean Opinion Score (DMOS) for the harsher noise conditions [5]. A Diagnostic Acceptability Measure (DAM) provided quality and diagnostic information.

Diagnostic Acceptability Measure (DAM)

Two conditions were tested using the DAM. The "Quiet" environment was recorded in an anechoic sound chamber using a dynamic microphone, and the "Office" environment was recorded in a modern office. Figure 1 shows these DAM scores.

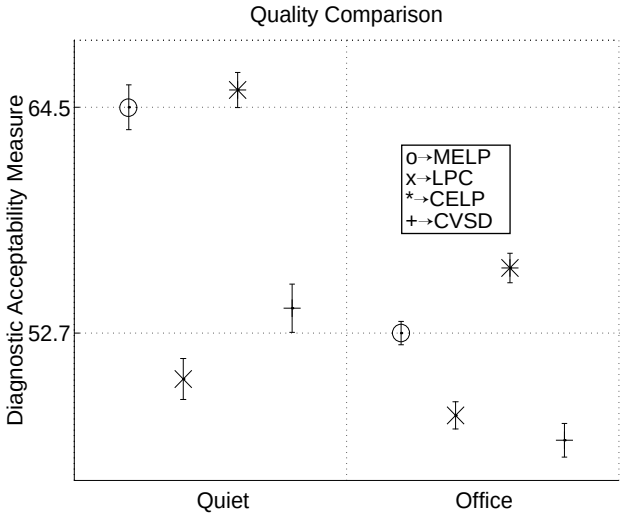

Figure 1: DAM Comparison

All coders performed better in the quiet environment than in the office environment. CELP outperformed all coders in both environments. MELP, however, performed only slightly lower than CELP, and well within the standard error for CELP in the quiet environment. LPC and CVSD alternated for lowest score in the two environments.

Male and Female DAM Scores

In nearly every case under both environments, female talkers scored higher than male talkers. The MELP coder was the only exception with the male talker score exceeding the female score under the quiet environment, but the difference was within the standard error. The difference between male and female scores in the quiet environment was always within the standard error, but the differences in the office environment were always greater than the standard error. The scores for female talkers exceeded those for the male talkers in the office environment by greater than five points for all coders except CVSD.

Mean Opinion Score

The MOS test included four acoustic noise conditions and two channel conditions. The "H250" environment was recorded in an anechoic sound chamber with an H250 Vinson microphone. Two error environments were also tested: a 1% random bit error channel and a 0.5% random block error channel. The block error contained 50% errors within a 35 millisecond block. "MCE" is a mobile command environment.

All coders scored MCE and 1% bit error conditions as the fifth and sixth ranked respectively. The quiet and 0.5% block errors

were ranked first and second for all but CVSD. Unexpectedly, quiet was ranked second for MELP and LPC behind the block error channel condition. Figure 2 shows MOS coder scores.

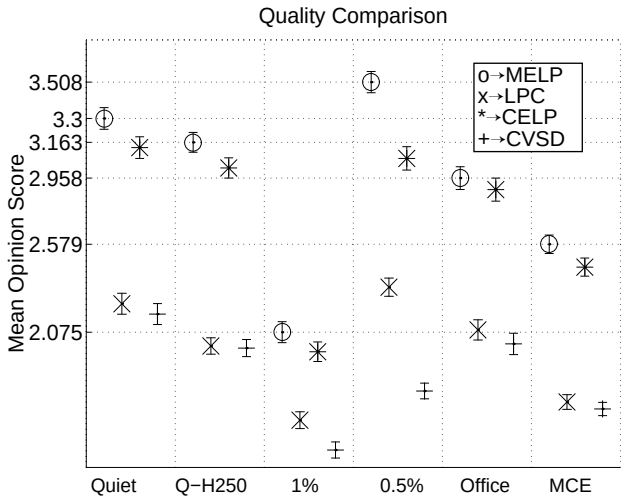

Figure 2: MOS Comparison

Relative coder ranking is easily seen in Figure 2. In all environments, MELP shows the highest MOS score, followed by CELP, LPC, and CVSD. MELP and CELP usually cluster together. LPC and CVSD also tend to cluster. These clustering trends do not apply to the block error channel environment. The Consortium test fixture [6] warned each coder when block errors were occurring: the older standard algorithm coders did not take advantage of this, and hence the MELP score is significantly elevated. In all cases there is a large separation between MELP-CELP scores and LPC-CVSD scores accurately indicating their differences in quality.

Male and Female MOS Scores

MELP and LPC coders both generally scored higher for male talkers than female talkers. Only in the 0.5% block error condition did the female MELP score exceed the male score, but this variance was within the standard error.

CELP and CVSD coders conversely both generally scored higher for female talkers than for male talkers. This was especially true in the office environment for both coders. CELP H250 microphone shaping also exhibited remarkably higher scores for female talkers.

Degradation Mean Opinion Score (DMOS)

The DMOS test included three acoustic noise conditions and two tandem conditions whose results are shown in Figure 3. The high mobility multi-wheeled vehicle (HMMWV) is considered the modern equivalent of the jeep. The E3A is the airborne warning and control system (AWACS) aircraft. The tested automobile was a Plymouth Reliant traveling on a highway. CVSD provided the basis for both tandem tests. In single tandem, the coder processed CVSD speech. In double tandem, CVSD processed the single tandem speech.

All coders except CVSD identically ranked the five conditions: double tandem, single tandem, E3A, auto, and HMMWV, from highest scoring to lowest scoring. The double tandem outranked the single tandem for all coders. As Figure 3 shows, either MELP or CVSD claimed the best score for each condition, while LPC is consistently ranked lowest.

Male and Female DMOS Scores

All of the coders scored higher for male talkers than for female talkers in nearly all of the environments. All of the coders except CVSD exhibited higher scores for female talkers in the single tandem environment. Single tandem was the only environment in which female talkers scored higher for the MELP coder. The CELP and LPC coders showed higher scores for females in the E3A environment.

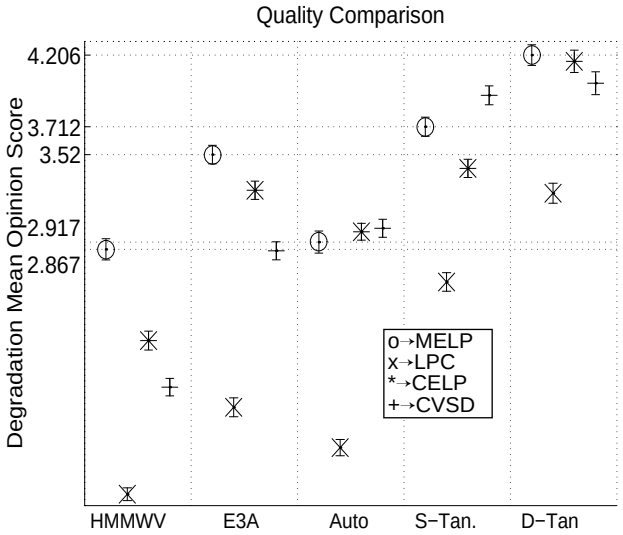

Figure 3: DMOS Comparison

INTELLIGIBILITY

Intelligibility testing was performed using the diagnostic rhyme test (DRT) [5]. Ten acoustic noise conditions, one microphone shaping condition, two channel noise conditions, and two tandem conditions were tested.

The same eleven conditions tested for quality were also tested for intelligibility. Four additional conditions were also added: the P3C Orion aircraft, the M2 Bradley tank, the F15 Eagle fighter aircraft, and the CH47 Chinook helicopter.

There is little similarity among the score rankings for each coder except that they all were challenged by the HMMWV and M2 environments. It's unusual to see a quiet score ranking lower than others, but the CVSD coder ranked the quiet environment fifth, below H250 microphone shaping, MCE, E3A, and office. Single tandem always provided higher intelligibility scores than double tandem. Three of the four coders (excepting LPC) consistently ranked F15, M2, CH47, and HMMWV as the four lowest scoring environments. These DRT scores are presented in Figures 4 and 5.

The environments included in Figure 4 are the more benign of the fifteen tested. In these environments MELP scored highest in all but P3C and MCE environments. Figure 5 shows all the harsher environment scores for all coders except LPC. LPC results for its three worse environments are M2: 38.4; HMMWV: 31.7; and CH47: 47.6. These scores are not shown in Figure 5.

MELP intelligibility does not compare as well in these harsher noise environments against CVSD. Only in F15 noise does MELP outperform these higher bit rate coders. In nearly all harsh conditions, CVSD exhibits the greatest intelligibility of all the coders. With the exception of the automobile, MELP outperforms CELP in every harsh environment. Both figures clearly show the persistent disparity between LPC and the other coders.

Male and Female DRT Scores

No coder exhibited many environments in which either male or female talkers displayed greater intelligibility. CELP and MELP slightly in favored male talkers; LPC and CVSD slightly favored female talkers. Female talkers scored higher for all four coders in HMMWV, CH47, and M2 environments. Conversely, male talkers scored higher for all coders in the single tandem, automobile, and P3C environments.

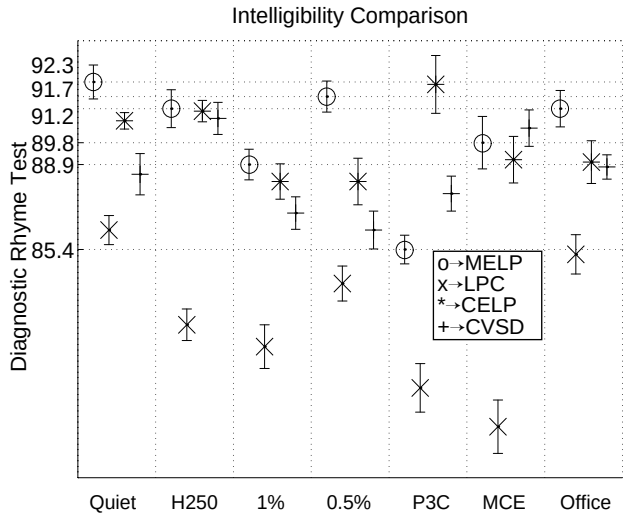

Figure 4: Intelligibility Comparison, Benign

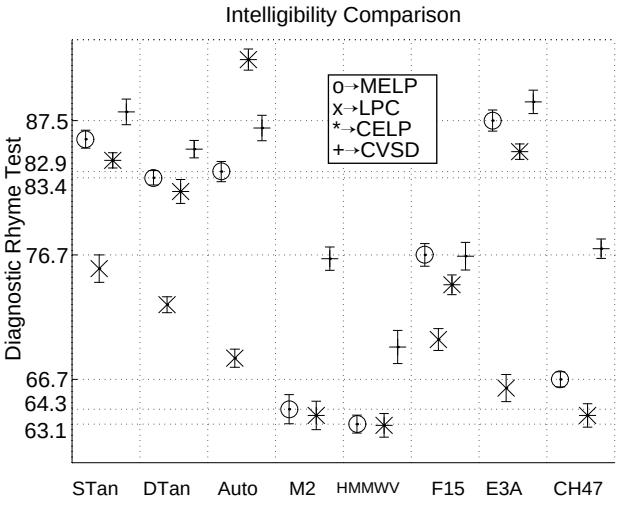

Figure 5: Intelligibility Comparison, Harsh

In most environments the difference between male and female scores exceeded the standard error, and in several the differences exceeded five DRT points. For MELP, scores for male talker scores greatly exceeded those for female talkers in the automobile environment, while the reverse was true in CH47 and M2 environments. For LPC, female scores exceeded male scores by more than fifteen DRT points in HMMWV and CH47 environments.

COMMUNICABILITY

Communicability testing was performed using the 1995 ARCON communicability exercise (ACE-95) [7]. Four different asymmetrical scenarios were tested thus yielding eight conditions. The D.o.D. scenario held communication between an office and quiet environment through a plain old telephone system (POTS) channel (Q/O and O/Q). The Army scenario held communication between a HMMWV and quiet environment across a satellite channel (H/Q and Q/H). The Air Force scenario held communication between an E3A and MCE across a joint tactical information distribution system (JTIDS) channel (E/M and M/E). The Navy scenario held communication between an aircraft carrier and an office across a high frequency (HF) channel (C/O and O/C).

No two coders yielded their best performance in the same scenario. MELP ranked the Quiet/Office scenario far higher than the others. CELP preferred the Office/Quiet version, while LPC and CVSD ranked the Army's two scenarios highest with little variance between them. The Navy scenarios were most commonly ranked as most difficult with the Carrier/Office path was ranking most difficult for all coders except LPC. Figure 6 shows a comparison of the communicability test across all coders.

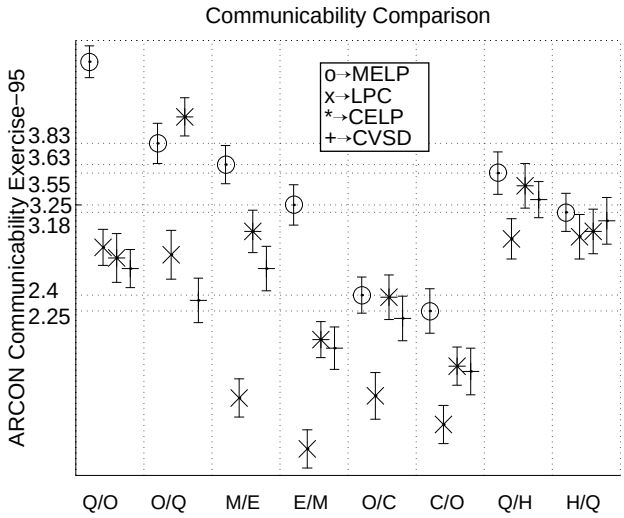

Figure 6: Communicability Comparison

RECOGNIZABILITY

Recognizability testing was performed using the Naval Research Laboratory (NRL) talker recognizability test (NTRT) [8]. Sentence pairs compared either processed voice with processed voice or unprocessed voice with to processed voice. All coders recognized talkers better when processed speech was compared with processed speech. Figure 7 provides these NTRT scores.

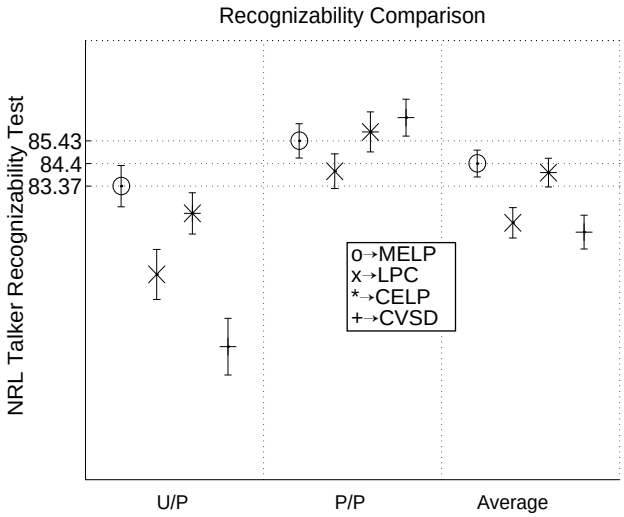

Figure 7: Recognizability Comparison

MELP excelled in the unprocessed versus processed comparison. CVSD, contrarily, shows a marked deficiency in this comparison but exhibited the highest score in the processed versus processed comparison. When averaging the two comparisons MELP obtained the highest score.

Male and Female NTRT Scores

In both comparisons for all coders, the male talkers scored higher than the female talkers. All differences between male and female scores well exceeded the standard errors.

COMPLEXITY

Complexity was measured using million instructions per second (MIPS), random access memory (RAM) and read only memory (ROM) measurements. MIPS were measured at the host lab using the real time equipment. Linker memory maps were analyzed to obtain the RAM and ROM measures.

As Table 1 shows, MELP complexity exceeds CVSD, LPC, and CELP in both processor and memory requirements. The MELP analyzer requires 72% of its total processing. These additional memory requirements are due to vector quantization tables which MELP uses for both line spectral frequencies (LSFs) and Fourier magnitudes.

Coder	MIPS	RAM	ROM
MELP	20.43	98.2k	128k
CELP	17.0	14.8k	128k
LPC	8.7	12.93k	128k
CVSD	0.1	1k	1k

Table 1: Complexity Comparison

DIAGNOSTICS

Two different tests provided diagnostic information on both the reference and the MELP coders.

Diagnostic Acceptability Measure Diagnostics

The ability of the DAM to provide diagnostic information, in the form of elementary perceptual qualities (EPQs), was the primary reason for performing this test. Eight signal EPQs and seven background EPQs are provided for each test. In all cases, the signal EPQs proved far more important toward overall degradation than were the background EPQs.

MELP shared no commonality with the other three standard coders in the EPQs contributing the most toward signal degradation. The “Babble” and “Nasal” EPQs contributed the most toward reducing the signal quality, with babble being the primary problem. These EPQs are affected by pitch, voicing, and peak clipping in addition to other factors.

All three of the other standard coders showed the “Interrupted” EPQ to be among the lowest scores. Interrupted speech is affected by dropouts in the speech. Other EPQs showing dominant scores for signal degradation for CELP, LPC, and CVSD were “Thin” and “Muffled”. Thin is best exemplified by high-passed speech; muffled, by low-pass speech.

Diagnostic Rhyme Test Diagnostics

Because the DRT measures the ability to distinguish between initial consonants, the diagnostic factors provided deal solely with consonants. The ability to distinguish between the presence or absence of a particular factor is measured in this test.

In nearly every condition for all four coders, “Graveness” was the factor contributing the most toward intelligibility degradation. In these few cases where graveness was not the primary contributor, it was almost always the secondary contributor.

Grave phonemes are produced by constriction toward the anterior of the vocal tract rather than the middle of the vocal tract, and they involve relatively steep upward transitions of the second formant. Grave phonemes can be either voiced or unvoiced, and the unvoiced can be either plosive or nonplosive. In all cases the unvoiced grave consonants were the source of the intelligibility decrease, and most of these were due to the nonplosive unvoiced grave phonemes.

Sustention was a source for intelligibility decrease in some environments for three of the coders. Sustained phonemes are produced by incomplete constriction of the vocal tract rather than complete constriction. These consonants are distinguished by their gradual onset and by the presence of mid-frequency noise. Sustained phonemes can be either voiced or unvoiced.

Both MELP and CELP exhibited problems with sustention in the HMMWV environment. Sustention was also a problem for CELP in the M2 and CH47 environments. Sustention errors were the primary source of intelligibility loss for LPC in the double-tandem and automobile environments. LPC and MELP exhibited more errors with voiced sustention, while CELP exhibited errors with both voiced and unvoiced sustention.

CONCLUSIONS

In general, the performance of MELP greatly exceeded that of LPC. The Consortium’s goal for CELP performance was satisfactorily achieved. In all but the harsher noise environments, the new coder also outperformed the CVSD coder.

While MELP outperforms LPC, CVSD, and CELP, its complexity is also correspondingly higher. But the target device for the new federal standard coder (based on user requirements) is an 80 MHz DSP device and four megabits of memory, and MELP demands only 51% of the target processor speed and 77% of the target memory.

REFERENCES

[1] L.M. Supplee, R.P. Cohn, J.S. Collura, A.V. McCree, “MELP: The New Federal Standard at 2400 bps”, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, 1997.

[2] M.A. Kohler, P. LaFollette, M. R. Bielefeld, “Criteria for the D.o.D. 2400 bps Vocoder Selection”, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, Georgia, USA, 1996

[3] J. P. Campbell, Jr., T. E. Tremain, and V. C. Welch. “The Federal Standard 1016 4800 bps CELP Voice Coder.” Digital Signal Processing 1, no. 3 (1991): 145 - 155.

[4] “Analog to Digital Conversion of Voice by 2400 bit/second Linear Predictive Coding”, Federal Standard 1015, Nov 1984.

[5] J.D. Tardelli, E.W. Kreamer, “Vocoder Intelligibility and Quality Test Methods”, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, Georgia, USA, 1996

[6] P.D. Gatewood, P.A. LaFollette, “Host Laboratory Functions for the DoD 2400 bps Vocoder Selection Process, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, Georgia, USA, 1996

[7] E.W. Kreamer and J.D. Tardelli, “Communicability Testing for Voice Coders”, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, Georgia, USA, 1996

[8] A.Schmidt-Nielsen, D.P.Brock, “Speaker Recognizability Testing For Voice Coders, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, Georgia, USA, 1996

[9] T.E. Tremain, M.A. Kohler, “Philosophy and Goals of the DoD 2400 bps Vocoder Selection Process, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, Georgia, USA, 1996

[10] M.R. Bielefeld, “1995 Test and Evaluation Plan for the Selection of a New U.S. Government Standard Voice Processor at 2,400 bps”, MITRE Technical Report 96W0000029, August 1996.

[11] V.C. Welch, T.E. Tremain, J.P. Campbell, Jr., “A Comparison of U.S. Government Standard Voice Coders”, *IEEE Military Communications Conference*, Boston, MA, 1989.

[12] W.D. Voiers, “Evaluating Processed Speech Using the Diagnostic Rhyme Test”, Speech Technology, Jan/Feb 1983.

[13] M.A. Kohler, “Analysis of Performance and Complexity for Four Government Standard Voice Coding Algorithm”, NSA Technical Report, R22-002-96, S-243,720.