NTIA Report 01-386

Voice Quality Assessment of Vocoders in Tandem Configuration

Christopher Redding
Nicholas DeMinco
Jeanne Lindner

U.S. DEPARTMENT OF COMMERCE
Donald L. Evans, Secretary

John F. Sopko, Acting Assistant Secretary
for Communications and Information

April 2001

This Page Intentionally Left Blank

This Page Intentionally Left Blank

PREFACE

This work was performed by the Institute for Telecommunication Sciences (ITS), Boulder, Colorado for the National Communications System's Office of Standards and Technology, Washington, DC under reimbursable order No. DNRO 66008.

The authors wish to thank Stephen Voran and Margaret Pinson of ITS for use of the APRE testing software used in the tests and for their expertise in the application of the software. Mr. Voran was a valuable resource for his expertise in voice quality of service measurements and vocoder theory of operation.

PRODUCT DISCLAIMER

Certain commercial equipment, components, and software are identified in this report to specify adequately the technical aspects of the reported results. In no case does such identification imply recommendation or endorsement by the National Telecommunications and Information Administration, nor does it imply that the equipment, components, or software identified is necessarily the best available for the particular application or uses.

CONTENTS

Page

FIGURES vi

TABLES vii

1. INTRODUCTION 1

2. SPEECH SIGNAL QUANTIZATION AND CODING 2
- 2.1 Speech Signal Quantization 2
- 2.2 Speech Signal Coding 3

3. TYPES OF VOCODERS TESTED 4
- 3.1 CELP Vocoder 5
- 3.2 VSELP Vocoder 5
- 3.3 QCELP Vocoder 6
- 3.4 IMBE Vocoder 6
- 3.5 AMBE Vocoder 7
- 3.6 ACELP Vocoder 7

4. MEASUREMENT STRATEGY AND DATA COLLECTION 7
- 4.1 Testing Software for Vocoder Speech Quality Estimation 8
- 4.2 Testing Techniques for Subjective Vocoder Performance Evaluation 10

5. TEST RESULTS 11

6. CONCLUSIONS 12

7. REFERENCES 18

APPENDIX A: MEASURED AND CALCULATED DATA 19

FIGURES

Page

Figure 1. Vocoders in tandem configuration 2

Figure 2. Test configuration 10

Figure 3. Mean L(AD) for tested vocoder configurations 14

Figure 4. Standard deviation of L(AD) for tested vocoder configurations 15

Figure 5. Estimated MOS for tested vocoder configurations 16

Figure 6. Estimated MOS standard deviation for tested vocoder configurations 17

TABLES

Page

Table 1. Vocoder Characteristics 4

Table 2. MOS Ranking and Quality Scale 11

Table A-1. Measured Data from APRE Test Set for L(AD) 19

Table A-2. Measured Data from APRE Test Set for Estimated MOS 20

Table A-3. Ninety-Five Percent Confidence Interval About the Mean for VSELP 21

This Page Intentionally Left Blank

This Page Intentionally Left Blank

VOICE QUALITY ASSESSMENT OF VOCODERS IN TANDEM CONFIGURATION

Christopher Redding, Nicholas DeMinco and Jeanne Lindner *

This report describes the objective testing of various vocoders, both singly and in different combinations of tandem configurations, to evaluate overall voice quality in an objective manner. The objective test evaluation was performed using a standardized (and patented) objective voice quality estimation algorithm developed by the Institute for Telecommunication Sciences (ITS). This method determines an auditory distance (AD) and converts it to a finite scale that can be related to the mean opinion score (MOS) scale commonly used in subjective voice quality tests. This testing establishes the feasibility (from a voice quality perspective) of using vocoders in a tandem configuration in order to relay voice communications between disparate systems.

Key words: Communications, vocoder, speech codec, tandem, voice quality of service, objective testing, logistic auditory distance, mean opinion score, subjective testing, perceived speech quality.

1. INTRODUCTION

Vocoders are used in digital voice communication systems to digitize and compress speech signals to minimize transmitted bit rate. Bandwidth is a precious commodity in wireless communication systems, since service providers must accommodate many users with a limited allocated bandwidth. Vocoders allow voice to be transmitted efficiently over circuit-switched or packet-switched digital networks. Vocoders also make spectrum-efficient wireless voice communications possible, and they allow for the digitized voice stream to be encrypted. It is a goal of vocoders to transmit the highest quality speech using the least amount of bandwidth. This must be accomplished using the lowest complexity to reduce cost of implementation and to reduce the amount of delay through the vocoder. More complex implementations result in more processing delay. There are many different types of vocoders designed to work with the many different types of communication systems. The variety of communication systems using different types of vocoders requires that there be decoding and encoding at systems interfaces if these different systems are required to provide voice communications to each other. Vocoders will be required for each digital system; and, in addition, the vocoders must be connected in a "tandem configuration" in order for communications to occur between two different systems (as shown in Figure 1). Manufacturers of vocoders generally focus their testing efforts on vocoders alone and spend less effort on the consideration of tandem configurations. It is the objective of this program to test the voice quality of various vocoders in different tandem configurations. The vocoder tests were performed first on individual vocoders and

---

* The authors are with the Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U.S. Department of Commerce, 325 Broadway, Boulder, CO 80305.

then on pairs of vocoders in a tandem configuration. The testing of different combinations of tandem configurations allows the performance of these vocoder combinations to be evaluated.

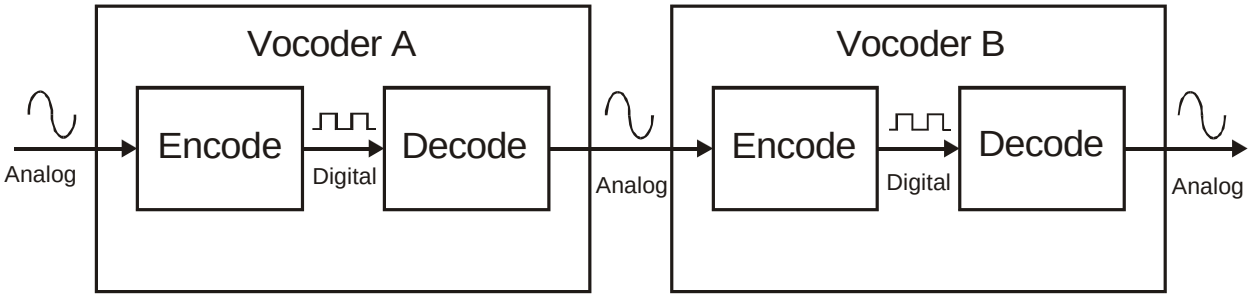

Figure 1. Vocoders in tandem configuration.

2. SPEECH SIGNAL QUANTIZATION AND CODING

2.1 Speech Signal Quantization

A vocoder contains an analog-to-digital (A/D) converter which samples the continuously varying amplitudes of the analog audio signal. The Nyquist theory for sampling a signal requires that the continuous analog signal be sampled at a rate of twice the highest frequency present in the spectrum of the sampled analog signal in order to accurately recreate the analog audio signal from the discrete samples. The analog audio signal must first be low-pass filtered to make it finite in bandwidth, which prevents any aliasing distortion from occurring. The vocoder then applies digital signal processing techniques to reduce the occupied bandwidth of the digital representation of the audio signal. This (A/D) conversion and bandwidth-reduction processing is referred to as encoding. The signal is then used to digitally modulate a carrier signal, which is then sent through a propagation channel, and demodulated at the receiver. The demodulated digital signal is then processed and converted back to analog audio using a decoder and a digital-to-analog (D/A) converter [1].

Since 4 kHz bandwidth speech is sufficient for communications purposes, Nyquist's sampling theory indicates that 8 kHz sampling is sufficient. If each sample is to be represented using b bits, then each sample must be mapped to one of 2^b discrete amplitude levels. This process is called quantization and it induces quantization errors. Quantization errors are essentially the round-off errors associated with replacing an arbitrary amplitude value with the nearest predetermined discrete amplitude level.

The bit-rate required by raw digitized speech is the product of the sample rate (in samples per second) and the quantization resolution (in bits per sample). For example, when speech is sampled at 8000 samples per second, and 16 bits are used to represent each sample, then the required bit rate is 128 kbps. Speech coders are often employed to reduce the bit-rate requirements.

2.2 Speech Signal Coding

Speech coders are classified into two categories based on how they accomplish signal compression. These two categories are waveform coders and vocoders.

2.2.1 Waveform Coders

A waveform coder reproduces the time waveform of the speech signal as closely as possible. Waveform coders code all signals equally well (speech and non-speech), since they do not tailor their coding techniques to unique characteristics of the signal. They are robust for all speech signal characteristics and have high immunity to noise, and in addition tend to have minimal complexity.

Waveform coders use simple algorithms that are inexpensive to implement and have short processing delays, since waveform coders are effectively sophisticated A/D converters. Because waveform coders do not compress the signal much, they do not conserve bandwidth as well as vocoders. Thus they require higher data rates to provide good fidelity. One example of a waveform coder is pulse code modulation (PCM). Uniform PCM uses uniformly spaced quantization levels. Other PCM formats such as μ-law or A-law PCM have quantization levels that are approximately logarithmically spaced, resulting in more quantization levels near zero amplitude and fewer quantization levels at large amplitudes. The result is an approximately constant signal-to-quantization noise ratio, which reduces the audibility of the quantization noise. This reduction means that more quantization noise can be tolerated, which translates into fewer bits per sample. Typically, only 8 bits per sample are required in μ-law or A-law PCM to attain the same speech quality of 12-14 bit per sample uniform PCM. This corresponds to a compression factor range of 0.57 to 0.67 (ratio of compressed bit-rate to original bit-rate).

The wired, circuit-switched telephone system in the United States primarily uses 8 bit per sample μ-law PCM and the wired, circuit-switched telephone system in Europe primarily uses A-law PCM. Both are defined in International Telecommunication Union (ITU) Recommendation G.711 [2]. G.711 PCM is found on T1, E1, and ISDN lines, it provides toll quality audio at 64 kbps, and it is often used as a quality benchmark against which other speech compression algorithms are judged.

Waveform coding techniques such as 64 kbps PCM can be replaced efficiently by speech coding techniques using vocoders. Vocoders make rates as low as 2.4 kbps [3] possible, but only with some trade-offs in terms of speech quality.

2.2.2 Vocoders

Vocoders are much more efficient users of available bandwidth than waveform coders, but tend to be more complex, with longer encoding and decoding delays, and lower speech quality. Vocoders are available in many different varieties, corresponding to the many possible trade-offs between complexity, bit-rate, speech quality, robustness to channel noise, and encoding-decoding delay.

Higher rate vocoders can provide speech quality that is nearly indistinguishable from PCM, while lower rate vocoders may show significant speech distortion and artifacts [1].

Vocoders typically process speech using 10-40 ms segments or frames. Bit rate reduction may be attained by exploiting the redundancy among the speech samples within a single frame and, to a much lesser extent, exploiting any redundancy of speech signal statistics between frames. Bit-rate reduction can also be attained through parameterized modeling of speech signal attributes such as pitch and power spectrum. Speech signal analysis at the encoder results in a parameter set that is passed to the decoder. The decoder uses these parameters to synthesize an approximation of the original speech signal.

3. TYPES OF VOCODERS TESTED

Several types of popular vocoders were investigated during this test. These included: the code excited linear predictor (CELP) vocoder specified in Federal Standard (FS) 1016, the vector sum excited linear predictor (VSELP) vocoder used in TIA/EIA/IS-54 time division multiple access (TDMA) systems, the Qualcomm code excited linear prediction (QCELP) vocoder used for TIA/EIA/IS-95 code division multiple access (CDMA) systems, the improved multi-band excited (IMBE) vocoder, the advanced multi-band excitation (AMBE) vocoder, and the algebraic codebook excited linear prediction (ACELP) vocoder used in the European terrestrial trunked (TETRA) radio system. All of these vocoders were hardware implementations, with the exception of the ACELP vocoder for TETRA which was a software-based implementation. Table 1 is a summary of the parameters and general information for the different vocoders used in the tests.

Table 1. Vocoder Characteristics

Vocoder	System	Implementation	Bit Rate (kbps)
CELP	Military	DSP[1]-based board	4.8
VSELP	TDMA	DSP-based board	8.0
QCELP	CDMA	Computer expansion board	9.6
IMBE	Project 25	DSP-based board	7.2
AMBE	Satellite	DSP-based board	4.8
ACELP	TETRA	Software	7.2

[1]Digital signal processing (DSP)

3.1 CELP Vocoder

The FS-1016 CELP vocoder has a bit rate of 4.8 kbps, a frame length of 30 ms, and it provides 4 kHz of voice bandwidth. It is used in many military communication systems, and it is also used for voice mail. It offers communications grade voice quality.

CELP stands for codebook excited linear prediction. Thus the core of CELP is linear predictive coding. In linear predictive coding, speech waveforms are analyzed in frames in order to determine how later samples in a frame could be predicted from earlier samples in that same frame. The resulting linear prediction coefficients provide a succinct description of the approximate relationship between the samples in the frame. There is a filtering interpretation of linear predictive coding as well. This interpretation views the linear prediction coefficients as a description of a filter that models the vocal tract transfer function associated with that frame of speech. We will use this filtering interpretation in the following description and refer to the filter as the vocal tract filter. A set of linear prediction coefficients (or an equivalent set of parameters) is generated by the encoder and sent to the decoder. The decoder can then use these coefficients to build the vocal tract filter. It remains to select a signal to excite the vocal tract filter. This can be done through further analysis at the encoder. The original speech signal is sent through the inverse of the vocal tract filter to determine the ideal excitation signal. If this ideal excitation signal were used to excite the vocal tract filter, the result would be an exact replica of the original speech signal. In practice, approximations to the ideal excitation signal are used in order to conserve bit rate.

In CELP, the ideal excitation signal is compared to a set of predetermined excitation signals (fixed and adaptive) stored in an excitation codebook. The codebook indices that lead to the best approximation of the ideal excitation signal are then transmitted from the encoder to the decoder. Since the decoder has a copy of the codebook, these received indices allow the decoder to construct the approximate excitation signal. When passed through the vocal tract filter, this approximation to the ideal excitation signal results in an approximation to the original speech signal. These processes are repeated for each new frame of speech.

3.2 VSELP Vocoder

The VSELP vocoder is used in EIA/TIA/IS-54-based TDMA systems. It has a bit rate of 8 kbps but is decreased to 7.95 kbps if the algorithm's in-band synchronization bit is not used. VSELP provides 4 kHz of voice bandwidth, has a frame length of 20 ms, and uses 5 ms subframes. It is used in the North American TDMA Cellular System Codec Specifications, which are continuing to evolve to achieve high quality speech using narrower RF bandwidth channels. The speech quality of VSELP might be characterized as either low toll grade quality or high communication grade quality.

The VSELP vocoder has modest computation complexity, is robust to channel errors, and uses three separate codebooks [5, 6]. The codebook output vectors are weighted and summed to generate an excitation sequence. This excitation sequence is used to update an adaptive codebook after each sub-frame.

3.3 QCELP Vocoder

The QCELP vocoder is a variable rate CELP vocoder that adjusts its output data rate according to specific characteristics of the input speech signal [4]. The range of output data rates is 800 to 9600 bps. The QCELP vocoder uses a frame length of 20 ms. This vocoder is used in all current generation CDMA (EIA/TIA/IS-95 or J-STD-008) spread spectrum systems. The implementation tested was a computer-based system evaluation board.

The goal of the variable rate vocoder is to minimize average bit-rate while holding speech quality approximately constant. Lower data rates are used for portions of the speech signal that are easily compressed, and higher rates are used for portions of the speech signal that are more complex. The natural pauses in speech which occur while listening, inhaling, or pausing between sentences are low energy or silent periods that are encoded at lower data rates [5, 6]. The active, high energy speech segments are encoded at high data rates. The energy thresholds used to make these determinations are dynamically adjusted to compensate for the level of background noise.

3.4 IMBE Vocoder

The IMBE and AMBE vocoders are based on work done at the Massachusetts Institute of Technology to develop a robust speech coding approach that would have potential advantages over the linear prediction approach.[2] Thus the IMBE and AMBE vocoders are not linear prediction based vocoders. In the linear prediction approach to speech coding, the choice of an excitation signal can model voiced speech (more harmonic) and unvoiced speech (more noise like) on a frame-by-frame basis. It follows that in the frequency domain, for each frame, the entire speech band is implicitly modeled as either voiced or unvoiced. In the multi-band excitation approach to speech coding, it is possible to partition the frequency band into noncontiguous voiced and unvoiced regions. This additional flexibility is equivalent to a more intricate model of speech and any background noise present. In fact, robustness to background noise is one of the stated motivations for pursuing the multiband excitation approach. Other arguments for the multiband excitation approach include the elimination of codebook searches (which can add significant complexity to CELP based vocoders) and ease of scalability to different bit rates.

The IMBE vocoder has been adopted for public safety use by the Association of Public Safety Communications Officials (APCO) Project 25 in North America and others. The total bit rate is 7.2 kbps (4.4 kbps for speech coding and 2.8 kbps for error correction coding), and the frame size is 20 ms.

The IMBE encoder extracts a set of parameters from the incoming speech signal including: pitch, a set of Voiced/Unvoiced (V/UV) parameters, and a set of spectral magnitudes. The V/UV analysis is based on the discrete Fourier transform (DFT) of a frame of speech samples, and is conducted across a number of frequency bands. Within each band, the V/UV parameters describe whether the

[2] http://www.dvsinc.com/products/software.htm, last accessed by ITS on July 7, 1999.

band contains periodic energy (voiced) or noise-like energy (unvoiced). The IMBE decoder uses this parameter set to reconstruct an approximation to the original speech signal. Filtered white noise is used to generate unvoiced components, and a bank of harmonic oscillators generate the voiced components. These two signals are combined to generate the final decoder output.

3.5 AMBE Vocoder

The AMBE vocoder is used in several satellite-based mobile communication systems. It has a 4.8 kbps sample rate. Its processing is much the same as the IMBE vocoder, but contains some algorithmic improvements that are said to lead to higher speech quality and greater robustness to both acoustic background noise and channel errors.

3.6 ACELP Vocoder

The ACELP vocoder considered here is the one used in the TETRA system as an open digital standard defined by the European Telecommunications Standards Institute (ETSI). The vocoder has a bit rate of 7.2 kbps including forward error correction (FEC). The ACELP vocoder algorithm is based on the CELP coding model, but ACELP codebooks have a specific algebraic structure imposed upon them. This structure is intended to allow for more efficient codebook searching and better fitting excitation signals. A frame length of 30 ms is used. These frames are further subdivided into 4 sub-frames that are 7.5 ms long. The linear prediction model used in ACELP includes a pair of filters. The first filter is a long-term prediction filter (pitch filter) which attempts to model the pseudo-periodicity in the speech signal. The second filter is a short-term prediction filter that models the speech spectral envelope [7].

4. MEASUREMENT STRATEGY AND DATA COLLECTION

The vocoders discussed in Section 3 were tested in different combinations of tandem configurations for voice quality of service (QOS). In tandem configuration, two vocoder pairs are connected end to end such that the audio output from the first vocoder is used as input to the second vocoder, as shown in Figure 1. ITS tested different combinations of tandem configurations using vocoders from different manufacturers in order to evaluate the degradation from these vocoder combinations.

For these tests, vocoders were tested individually as well as in tandem configurations. In tandem configuration the order of the vocoders was reversed so that each vocoder was operated in both the first and the second position of the tandem. The CELP vocoder was not tested in tandem but was included as a comparison for other single vocoders. The IMBE and AMBE vocoders were implemented on evaluation boards with 96 pin DIN 41612 connectors. The jumpers on the boards were configured to place the boards in a loop-back mode, where the analog audio input is immediately converted to a digital signal and then encoded to reduce bandwidth and then decoded and converted back to an analog signal. The IMBE and the AMBE boards had the same pin configuration; therefore, the same setup could be used on a custom-built test configuration.

Both the VSELP and CELP vocoders were implemented on a DSP board. The QCELP and 64 kbps PCM were implemented on the QCELP system evaluation boards. The QCELP system included two boards, which were installed into the PC bus expansion slots of two separate personal computers. This system could not be connected in a direct loop-back mode, so a vocoder board in one computer encoded the signal and a second board in the other computer decoded the signal. A serial data cable connected the two computers to provide a path for the digitally encoded signal. The ACELP vocoder was implemented in software.

Cables with audio connectors were usually used to interconnect the vocoders and test equipment; however, the ACELP vocoder was implemented in a software configuration on two computers. As a result, the file names needed to be manipulated so that speech that had been through the ACELP vocoder could be used as input to another vocoder. Speech files could also be sent through other vocoders and then played through an ACELP file. A software program, described in the next section, was used to perform the objective testing.

4.1 Testing Software for Vocoder Speech Quality Estimation

The Audio Play, Record, and Estimate (APRE) software tool was developed at ITS. APRE can be used to play a digital speech file into a device under test (DUT) while simultaneously making digital recordings of both the DUT input signal and the DUT output signal. These digital recordings are stored in files on a PC. The APRE software can then be used to apply a Measuring Normalizing Block (MNB) [10-12] algorithm that estimates the auditory distance (AD) between the DUT input and output files. Note that for these tests the DUT was either a single vocoder or two vocoders connected in a tandem configuration, with the output of the first vocoder being used as input to the second device.

The speech material used in these tests was recorded by two male and two female English language speakers. The sentences recorded came from the 1965 revised list of phonetically balanced sentences (usually referred to as the Harvard Sentences). This list is contained in reference [8] which is also IEEE Standards Publication No. 297. This list provides a good mix of English language phonemes for a balanced source of speech. A series of 40 sentences were played through the vocoder DUT. Half of these sentences were spoken by a male and half of the sentences were spoken by a female.

The APRE software was used to play these sentences into the DUT (Vocoder A) while simultaneously making digital recordings of both the DUT input signal and the DUT (Vocoder B) output signal, as shown in Figure 2. This is done using a standard sound card in a personal computer. One channel of the sound card line level output is connected to the DUT input and also looped back to the left channel of the sound card line level input. The output of the DUT is then connected to the right channel of the line level input on the sound card. Thus a stereo recording process results in synchronized DUT input and output signals recorded as the left and right signals of a stereo pair. It is these two recordings that are further analyzed by the APRE software. The APRE software first must estimate the delay between these two signals and compensate for it. The two signals, with delay removed, are then presented to an MNB algorithm. APRE offers a choice

of four different MNB algorithms. The tests described in this report use the version that has been standardized in ITU-T Recommendation P.861 [9]. This MNB algorithm calculates an AD between these two signals. If the DUT were transparent, the two signals would be identical, and AD would be zero. As the DUT changes the signal in more and more perceptually significant ways (e.g. more and more distortion) the AD between the two signals increases. Thus we see that AD values can range (in theory at least) from 0 to infinity.

The MNB algorithm calculates AD using a simple perceptual transformation (or model for hearing) and a more sophisticated distance measure (or model for judgment) [10-12]. The perceptual transformation includes mapping from a Hertz frequency scale to a more perceptually relevant Bark or Critical Band frequency scale. It also applies a compressive nonlinearity to signal amplitudes in order to map them to a more perceptually relevant signal loudness domain.

The distance measure is comprised of a hierarchical structure of MNBs. This structure emulates the ability of a listener to adapt and react to spectral differences. Time measuring normalizing blocks (TMNBs) and frequency measuring normalizing blocks (FMNBs) take perceptually transformed input and output speech signals as inputs and generate a set of measurements and a normalized version of the original output speech signal. A TMNB first integrates over some frequency scale and then measures differences and normalizes the output signal at multiple times. A set of FMNBs integrate over some time scale, then measure the difference and normalize the output signal at multiple frequencies. This hierarchical structure of MNBs allows for the measuring of spectral deviations at multiple time and frequency scales. This structure starts with larger time and frequency scales and works down to smaller scales, since this approach is most likely to emulate listeners' adaptation and reaction to spectral differences. The results from linearly independent MNBs are then linearly combined to generate AD:

$$AD = w^T \bullet m$$

where m is a vector of MNB measurements and w is a vector of weights.

AD can then be mapped into a finite range from one to zero using the logistic function:

$$L(AD) = \frac{1}{1 + e^{aAD+b}}$$

The constant a is positive and b is a negative number, so this logistic function has asymptotes at 0 (as AD gets large) and 1 (as AD approaches zero). Thus DUTs with small distortions will have L(AD) values near 1, while DUTs with large distortions will have L(AD) values near 0. L(AD) is the parameter that is used to compare the various vocoder configurations in this report. The L(AD) values are estimates of perceived speech quality. It has been demonstrated that these estimates do show good correlation to subjective test results on the mean opinion score (MOS) scale for a wide

variety of conditions [11,12], but of course they are not replacements for formal subjective tests. To move the L(AD) scale to cover the same interval as the MOS scale, we apply the following transformation:

$$MOS = 4L(AD) + 1$$

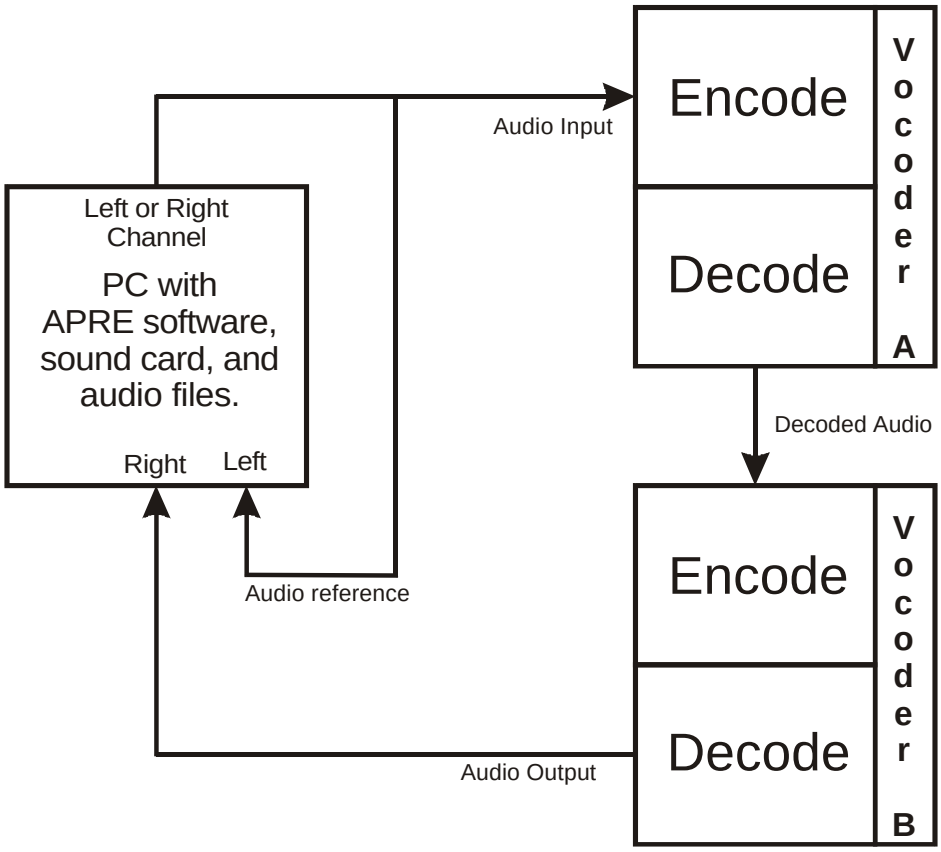

Figure 2. Test configuration.

4.2 Testing Techniques for Subjective Vocoder Performance Evaluation

Subjective voice quality testing has become increasingly important because of the popularity of wireless telephony and the development of new speech codecs (vocoders) for better speech compression and hence better spectrum utilization of the crowded bands of the spectrum. The commonly accepted opinion is that voice quality is a subjective parameter and the determination of voice quality should be performed using the results of subjective testing and applying the concept of MOS. However, the objective testing described in the previous section can save time in testing audio devices when compared to subjective testing [10,11].

In MOS testing, a panel of listeners listens to and rates the quality of speech samples that have been recorded through the DUT of interest. The listeners are asked to rate the samples of processed audio material on a five-point scale that represents their opinion of the quality of the audio sample. These five points are defined in Table 2 below. Table 2 describes the five levels of MOS [13,14]. The ratings from each listening case are averaged together and the result is the MOS. The listeners listen to these audio samples over headsets, handsets, or loudspeakers, depending on the particular test constraints or criteria.

Table 2. MOS Ranking and Quality Scale

Score	Quality Scale
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Because MOS is not an absolute or fixed number, one must be careful in the application or final use of MOS. The use and interpretation of MOS must take into account the situation or context in which the information was collected. The range of quality of the gathered samples has a direct bearing on the interpretation of the data. The absolute number for any particular case in the study is not as important as the differences in the numbers for the test cases in the study, because the test data were gathered under the same conditions, and relative quality ratings are valid methods of comparison. It must also be kept in mind that there are a large number of variables in a system simulation that cannot always be accurately characterized. The entries in this table are for toll quality in hard-wired telephone systems, which users generally have higher expectations for than that for wireless communication systems [13].

5. TEST RESULTS

The specific vocoders tested and the test results for vocoders and tandem combinations of vocoders are shown in Figures 3 - 6. The mean L(AD) for each of the vocoders and the tandem combinations of the vocoders are shown in Figure 3. The first seven bars are the single vocoders and one waveform coder (i.e., the 64 kbps PCM) that were tested. The single IMBE vocoder has a mean L(AD) of approximately 0.58. The other vocoders are similar to this in performance. The 64 kbps PCM waveform coder performs much better, almost approaching the ideal performance of a straight through wire, but it requires a 64 kbps data rate and the necessary bandwidth for transmission of that data rate. The eighth bar in this figure, labeled “Bypass,” is the result that would occur for a straight through connection with no degradation of the voice signal. This situation is the best case scenario. The first vocoder shown for the tandem listings is the input vocoder and the second is the output vocoder. Measurements were made for both combinations. These tests implicitly assume error-free

transmission channels so propagation effects in the channels are not included in this test. Thus, the performance measured is the best one can expect when the respective vocoder pairs are operated in tandem. When each vocoder is tested in a tandem configuration with 64 kbps PCM, the resulting performance is equivalent to or less than the performance of the vocoder, because a tandem configuration generally can be no better than the poorest performing vocoder in the tandem. Figure 4 shows the standard deviation of the L(AD) calculated across the 40 sentences used in these tests. This figure indicates that the variability of the L(AD) is small in comparison to the magnitude of the L(AD).

MOS is the most popular method for ranking vocoder performance when the tests are performed in a subjective manner. As described previously, there is an approximate mathematical relationship between the L(AD) and the MOS subjective results. This comparison is appropriate to use as long as it is recognized that this is a comparison of objective and subjective test results. Figures 5 and 6 are the estimated MOS means and standard deviations for the vocoders and combinations of vocoders. These values were calculated from the equations in Section 4.1. The estimated MOS for the single IMBE vocoder is approximately 3.4 and again represents acceptable quality for a mobile communication system. Tables A-1 and A-2 in the Appendix contain a summary of the actual numbers calculated by APRE for the test results shown in Figures 3 through 6. Both the mean and standard deviation, in addition to the upper and lower bounds for the 95% confidence interval (in the last 2 columns of Tables A-1 and A-2), are given in Tables A-1 and A-2 for the L(AD) and the MOS.

6. CONCLUSIONS

The test results show that the vocoder speech quality degrades when tandem configurations are used, as compared to a single vocoder. Comparison of the amount of degradation can be done by comparing each individual vocoder to that of various tandem configurations, shown by the mean L(AD) and estimated MOS scores in Figures 3 and 5. The single vocoders have nearly equivalent estimated MOS which ranges between 3 and 3.5. In tandem configuration, not including the PCM tandems, the MOS ranges between 2.5 and 3. For PCM tandems the MOS scores fall between 3 and 3.6. The MOS is decreased by 0.5 when tandem configurations are used as opposed to single vocoders, a decrease in voice quality that is not generally discernable by the untrained listener. The PCM tandems showed less degradation which is not unexpected due to the higher bit rate of the PCM coder. In general, tandem configurations degrade the QOS and will increase the end-to-end propagation delay times as compared to a single vocoder. The results presented in this report will help system designers, integrators, and users by showing the best voice quality that can be expected when interconnecting vocoders, because these tests were conducted under ideal conditions.

Performance criteria consisted of L(AD) with corresponding estimated MOS obtained from transformed L(AD). The estimated MOS values presented in this report are only estimates of perceived speech quality. They are provided here to give general guidance on coarse, relative, quality levels. It has been demonstrated that these estimates do show good correlation to subjective test results on the MOS scale for a wide variety of conditions [11,12]. However these estimates are not

intended as replacements for formal subjective tests. Further, it is possible that some of the DUT configurations presented here may contain distortions that would fall outside the scope of the evaluations described in [11,12], and little is known about the behavior of the estimates for any such DUT configurations.

One unexpected result deserves comment here. Compared to most vocoders, PCM is a waveform coder that is nearly transparent. Therefore, it is not surprising that the VSELP-PCM tandem and the VSELP have mean L(AD) values that are statistically equivalent (see Table A-3). However, the VSELP-PCM tandem has a mean L(AD) value that is slightly higher than the VSELP mean L(AD) value. Increased L(AD) values are similarly observed for all of the PCM tandems. This difference is just barely statistically significant (see Table A-3). Since higher L(AD) values correspond to higher estimated MOS values, this result is counter-intuitive. The rationale for this occurrence could be that the bandpass filtering associated with the PCM implementation creates a speech signal that is slightly easier for the VSELP vocoder to code with good fidelity. It may also be that this is a shortcoming of the objective estimator L(AD).

As noted, the vocoder systems tested had relatively similar estimated MOS values. The measurements were all performed using simulated error-free channels. A question arises as to what the performance would be in degraded propagation channels with multipath, noise, and other environmental effects superimposed on the propagation channels. Further vocoder testing using degraded channel conditions would need to be performed to quantify performance in this environment. These tests could be performed quicky and inexpensively using the objective test evaluation methods described in this report. On the other hand, equivalent subjective testing would be time and cost intensive. References 10 and 11 describe the results of comparisons of using the objective method using L(AD) and six other established estimators of perceived speech quality with the results of formal subjective tests. When compared to other methods described in References 10 and 11 of evaluating speech quality of vocoders, the method of using L(AD) with MNB estimators provided the best results across all conditions considered in the tests, and this method also has an improved ability to estimate perceived speech quality for lower rate vocoders [11].

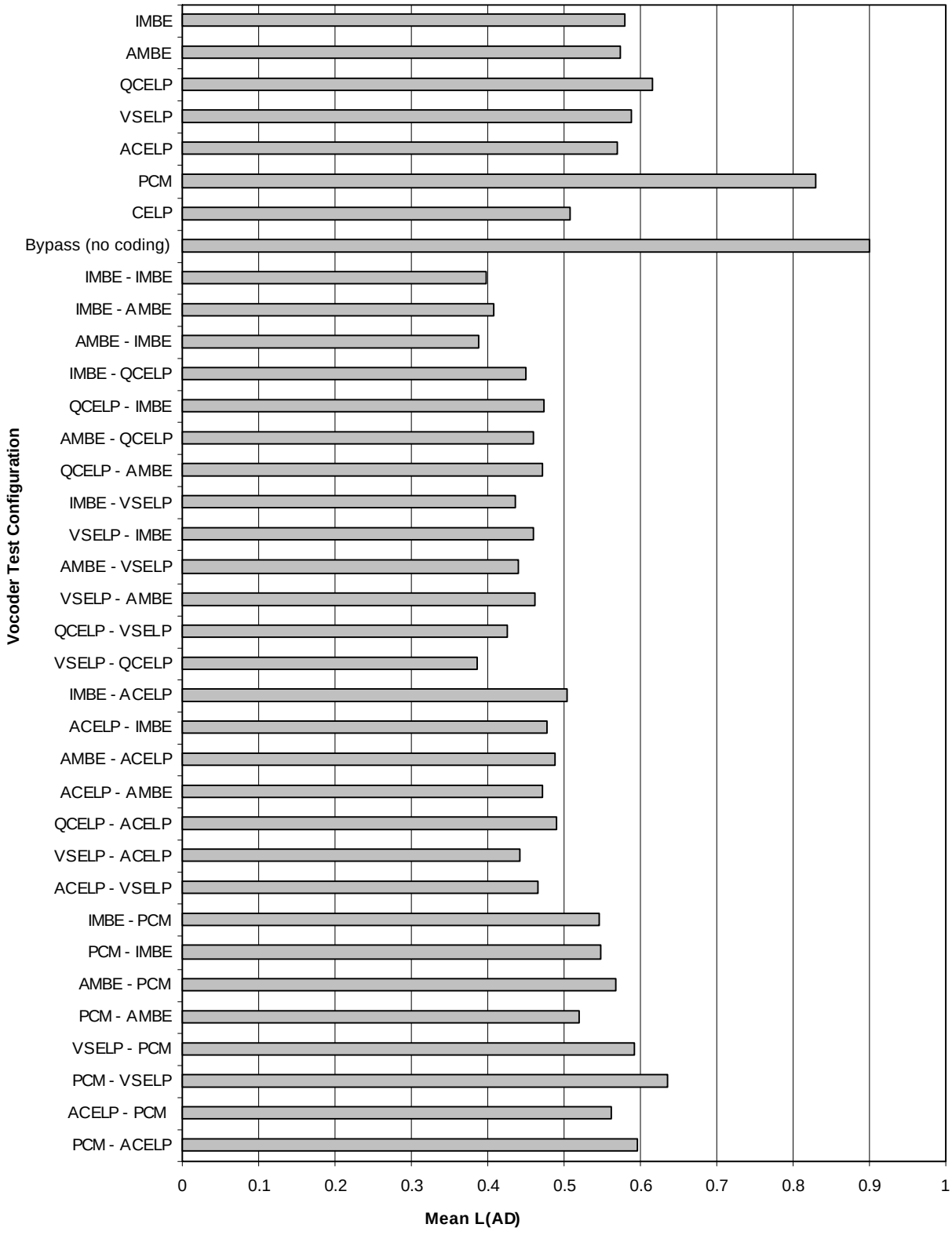

Figure 3. Mean L(AD) for tested vocoder configurations.

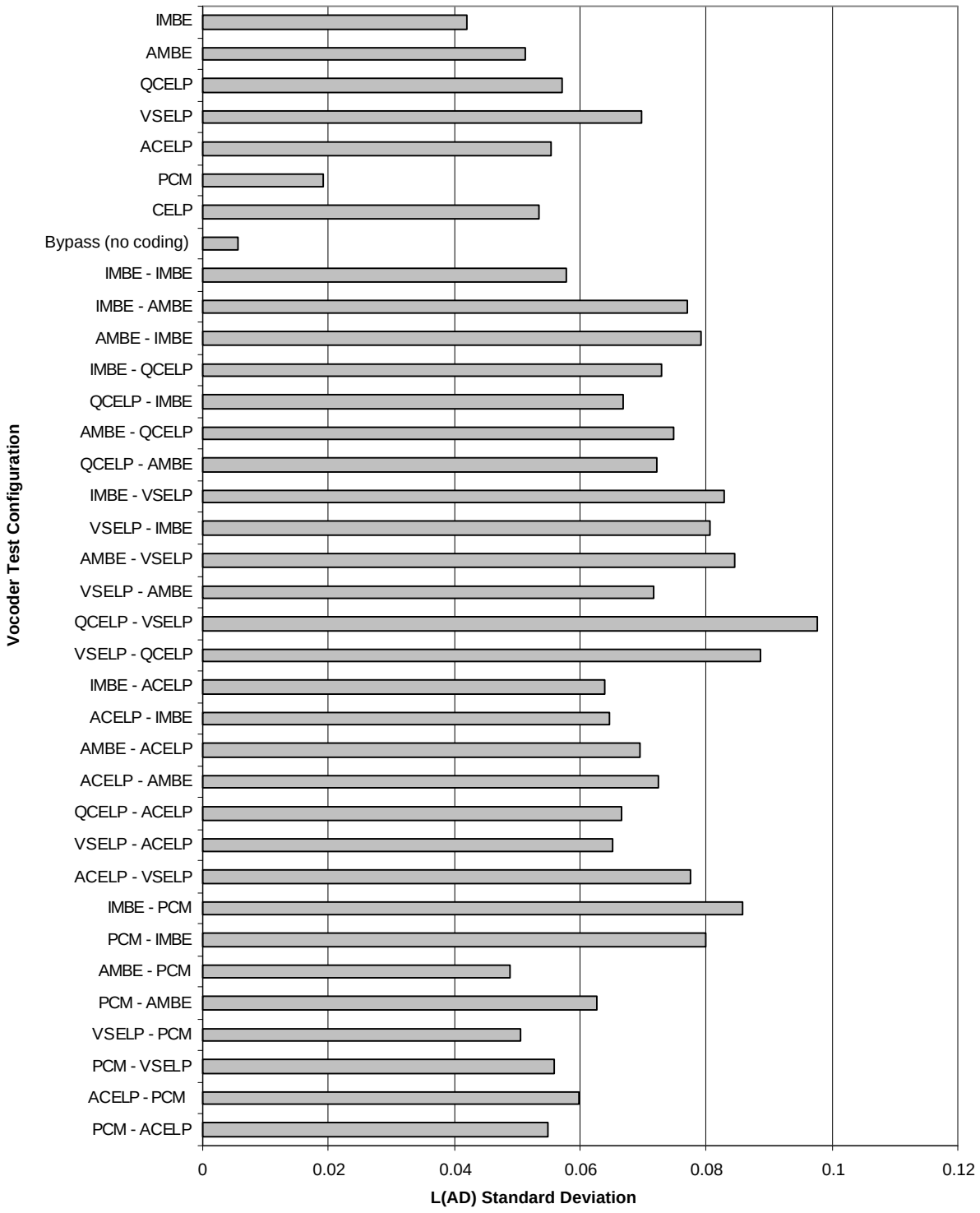

Figure 4. Standard deviation of L(AD) for tested vocoder configurations.

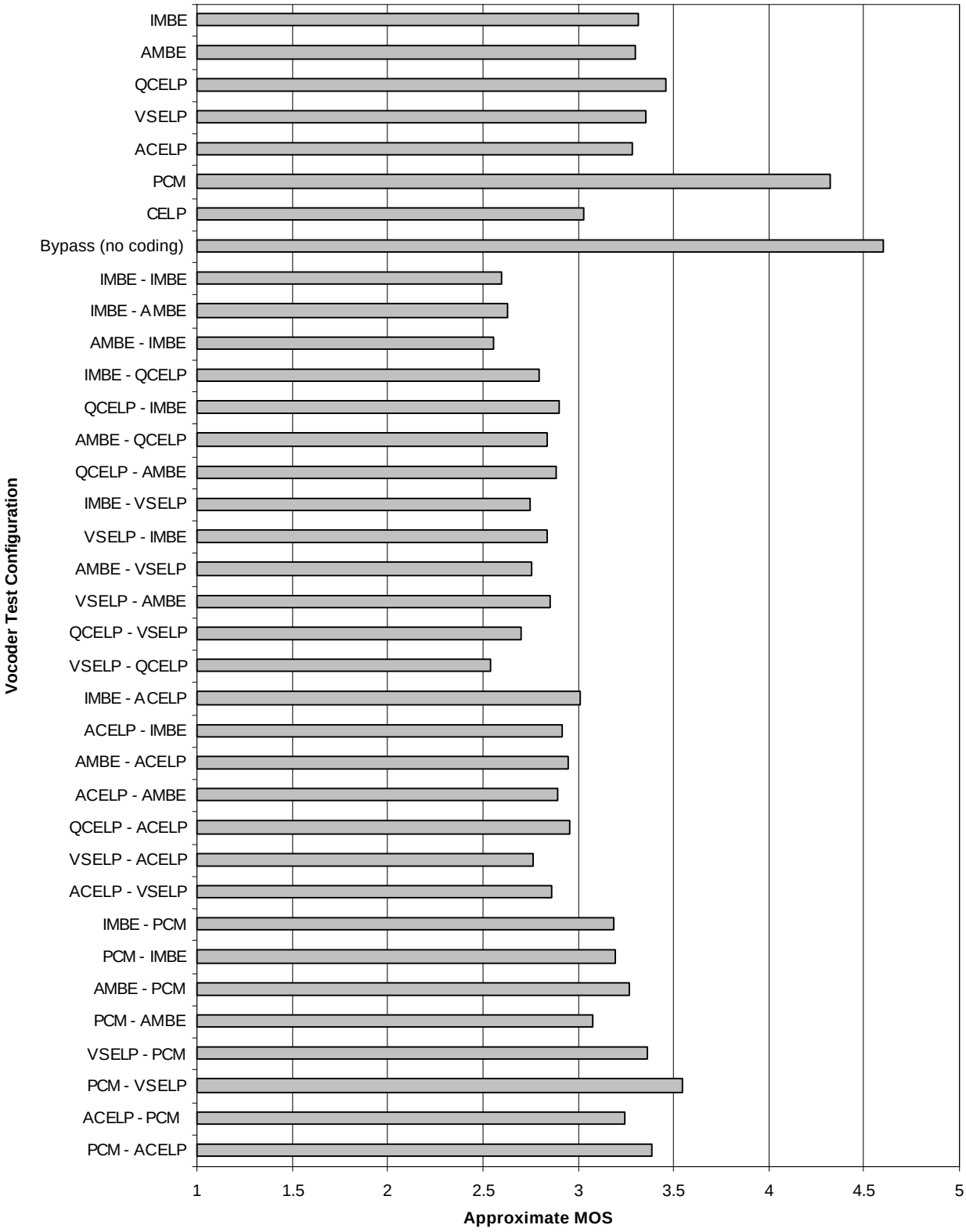

Figure 5. Estimated MOS for tested vocoder configurations.

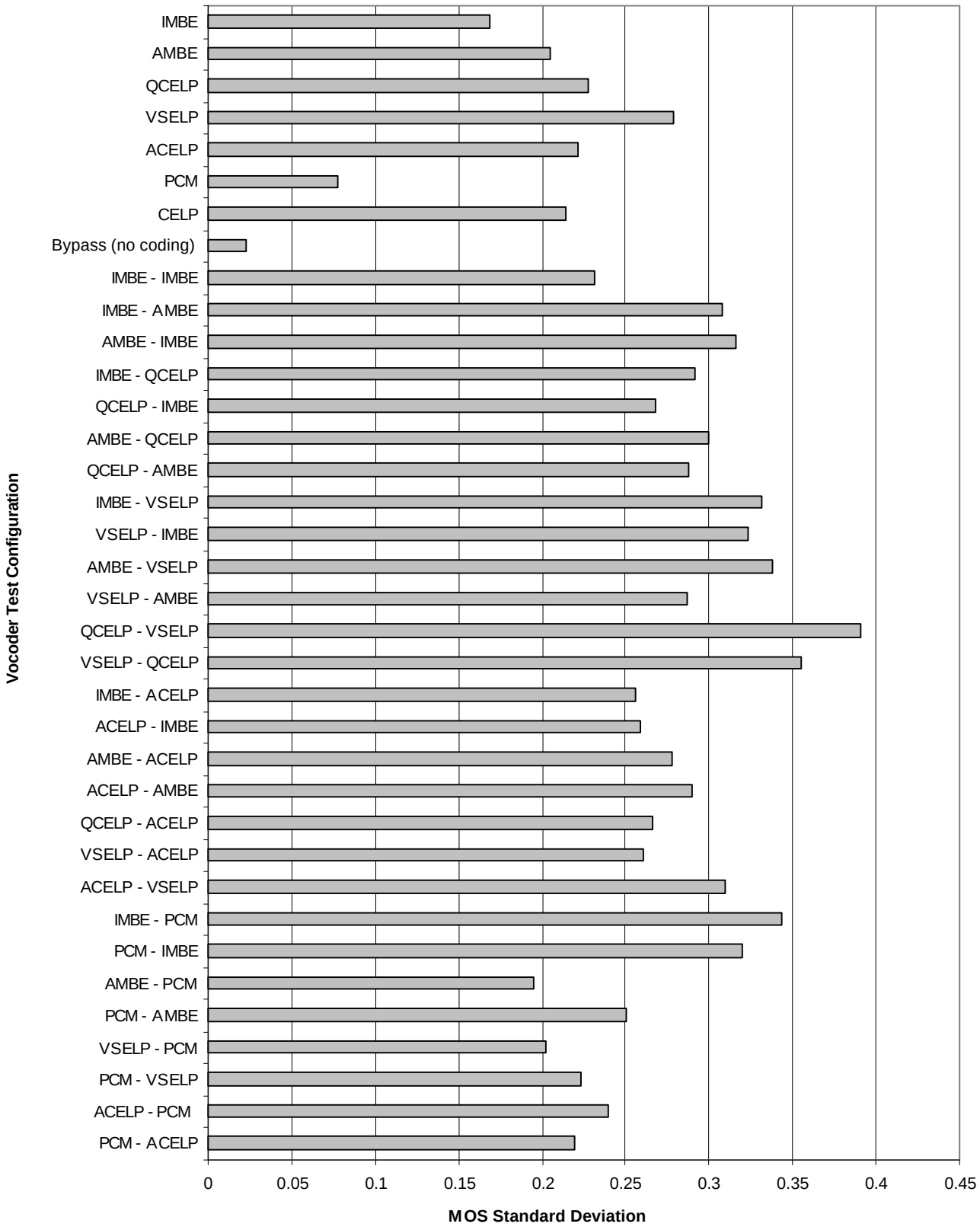

Figure 6. Estimated MOS standard deviation for tested vocoder configurations.

7. REFERENCES

[1] A.S. Spanias, "Speech coding: A tutorial review," *Proc. of the IEEE*, Vol. 82, No. 10, October 1994, pp.1541-1582.

[2] ITU Recommendation G.711, "Pulse Code Modulation (PCM) of Voice Frequencies," Geneva, 1988.

[3] L.M. Supplee, "MELP: The new standard at 2400 BPS," in *Proc. ICASSP '97*, 1997, pp. 1591-1594.

[4] *Vocoder Data Book*, Qualcomm Incorporated staff members, San Diego, CA, 1999.

[5] T.S. Rappaport, *Wireless Communications*, New Jersey: Prentice Hall, 1996, pp. 361-392.

[6] J. Bellamy, *Digital Telephony*, New York: John Wiley & Sons, Inc., 1991, pp. 93-159.

[7] European Telecommunication Standard Institute (ETSI), ETSI 300-395-1, May 1997, "Terrestrial Trunked Radio (TETRA); Speech CODEC for full-rate traffic channel; Part 1: General Description of Speech Functions."

[8] "IEEE Recommended Practice for Speech Quality Measurements," *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-17, No. 3, pp. 225-246, Sept. 1969.

[9] ITU Recommendation P.861, Appendix II, 1998 and ANSI T1.518, 1998.

[10] S.D. Voran, "Objective estimation of perceived speech quality- Part I: Development of the measuring normalizing block technique," *IEEE Trans. Speech and Audio Processing*, Vol. 7, No. 4, pp. 371-382, July 1999.

[11] S.D. Voran, "Objective estimation of perceived speech quality- Part II: Evaluation of the measuring normalizing block technique," *IEEE Trans. Speech and Audio Processing*, Vol. 7, No. 4, pp. 383-390, July 1999.

[12] S.D. Voran, "Objective estimation of perceived speech quality using measuring normalizing blocks," NTIA Report 98-347, Apr. 1998.

[13] ITU-T Recommendation P.800, Appendix B, "Methods for Subjective Determination of Transmission Quality," Geneva, 1996.

[14] A. Coleman, et al., "Subjective performance evaluation of the REP-LTP codec for the Pan European cellular digital mobile radio system," in *Proceedings of ICASSP*, pp. 1075-1079, 1989.

APPENDIX A: MEASURED AND CALCULATED DATA

Table A-1. Measured Data from APRE Test Set for L(AD)

Vocoder Configuration	Mean L(AD)	L(AD) Standard Deviation	L(AD) 95% Confidence Interval Lower Bound	L(AD) 95% Confidence Interval Upper Bound
IMBE	0.579	0.042	0.566	0.592
AMBE	0.574	0.051	0.558	0.590
QCELP	0.615	0.057	0.597	0.633
VSELP	0.588	0.070	0.566	0.610
ACELP	0.570	0.055	0.533	0.587
PCM	0.830	0.019	0.824	0.836
CELP	0.508	0.053	0.490	0.526
Bypass (no coding)	0.900	0.006	0.898	0.902
IMBE - IMBE	0.394	0.058	0.376	0.412
IMBE - AMBE	0.408	0.077	0.384	0.432
AMBE - IMBE	0.389	0.079	0.365	0.413
IMBE - QCELP	0.450	0.073	0.427	0.473
QCELP - IMBE	0.474	0.067	0.453	0.495
AMBE - QCELP	0.459	0.075	0.436	0.482
QCELP - AMBE	0.471	0.072	0.449	0.493
IMBE - VSELP	0.437	0.083	0.411	0.463
VSELP - IMBE	0.459	0.081	0.434	0.484
AMBE - VSELP	0.439	0.084	0.413	0.465
VSELP - AMBE	0.462	0.072	0.440	0.484
QCELP - VSELP	0.426	0.098	0.396	0.456
VSELP - QCELP	0.385	0.089	0.357	0.413
IMBE - ACELP	0.503	0.064	0.483	0.523
ACELP - IMBE	0.479	0.065	0.459	0.499
AMBE - ACELP	0.488	0.070	0.466	0.510
ACELP - AMBE	0.473	0.072	0.451	0.495
QCELP - ACELP	0.490	0.066	0.470	0.510
VSELP - ACELP	0.441	0.065	0.421	0.461
ACELP - VSELP	0.466	0.077	0.442	0.490
IMBE - PCM	0.547	0.086	0.520	0.574
PCM - IMBE	0.549	0.080	0.524	0.574
AMBE - PCM	0.568	0.049	0.553	0.583
PCM - AMBE	0.519	0.063	0.499	0.539
VSELP - PCM	0.592	0.051	0.576	0.607
PCM - VSELP	0.636	0.056	0.619	0.653
ACELP - PCM	0.562	0.060	0.543	0.581
PCM - ACELP	0.597	0.055	0.580	0.614

Table A-2. Measured Data from APRE Test Set for Estimated MOS

Vocoder Configuration	Estimated MOS Mean	Estimated MOS Standard Deviation	Estimated MOS 95% Confidence Interval Lower Bound	Estimated MOS 95% Confidence Interval Upper Bound
IMBE	3.32	0.168	3.27	3.37
AMBE	3.30	0.205	3.24	3.36
QCELP	3.46	0.228	3.39	3.53
VSELP	3.35	0.279	3.26	3.44
ACELP	3.28	0.222	3.21	3.35
PCM	4.32	0.077	4.30	4.34
CELP	3.03	0.214	2.96	3.10
Bypass (no coding)	4.60	0.023	4.59	4.61
IMBE - IMBE	2.59	0.231	2.52	2.66
IMBE - AMBE	2.63	0.308	2.54	2.73
AMBE - IMBE	2.56	0.317	2.46	2.66
IMBE - QCELP	2.80	0.292	2.71	2.89
QCELP - IMBE	2.90	0.268	2.82	2.98
AMBE - QCELP	2.84	0.299	2.75	2.93
QCELP - AMBE	2.88	0.288	2.79	2.97
IMBE - VSELP	2.75	0.331	2.65	2.85
VSELP - IMBE	2.84	0.323	2.74	2.94
AMBE - VSELP	2.76	0.338	2.66	2.87
VSELP - AMBE	2.85	0.287	2.76	2.94
QCELP - VSELP	2.70	0.391	2.58	2.82
VSELP - QCELP	2.54	0.355	2.43	2.65
IMBE - ACELP	3.01	0.256	2.93	3.09
ACELP - IMBE	2.91	0.258	2.83	2.99
AMBE - ACELP	2.95	0.278	2.86	3.04
ACELP - AMBE	2.89	0.290	2.80	2.98
QCELP - ACELP	2.96	0.266	2.88	3.04
VSELP - ACELP	2.76	0.260	2.68	2.84
ACELP - VSELP	2.86	0.310	2.76	2.96
IMBE - PCM	3.19	0.343	3.08	3.30
PCM - IMBE	3.19	0.320	3.09	3.29
AMBE - PCM	3.27	0.195	3.21	3.33
PCM - AMBE	3.08	0.251	3.00	3.16
VSELP - PCM	3.37	0.203	3.31	3.43
PCM - VSELP	3.54	0.223	3.47	3.61
ACELP - PCM	3.25	0.239	3.18	3.32
PCM - ACELP	3.39	0.219	3.32	3.46

Table A-3. Ninety-Five Percent Confidence Interval About the Mean for VSELP

Vocoder Configuration	Mean L(AD)	L(AD) 95% Confidence Interval Lower Bound	L(AD) 95% Confidence Interval Upper Bound	Estimated MOS Mean	Estimated MOS 95% Confidence Interval Lower Bound	Estimated MOS 95% Confidence Interval Upper Bound
VSELP	0.588	0.566	0.610	3.35	3.26	3.44
VSELP - PCM	0.591	0.592	0.608	3.37	3.31	3.43
PCM - VSELP	0.636	0.619	0.653	3.54	3.47	3.61