Audio Engineering Society

Convention Paper

Presented at the 123rd Convention
2007 October 5–8 New York, NY, USA

The Effects of Latency on Live Sound Monitoring

Michael Lester[1] and Jon Boley[2]

[1] Purdue University, Indianapolis, IN 46202, USA
(Now at Shure Incorporated, Niles, IL 60714, USA)
lester_michael@shure.com

[2] Shure Incorporated, Niles, IL 60714, USA
(Now at Purdue University, West Lafayette, IN 47907, USA)
jdboley@purdue.edu

A subjective listening test was conducted to determine how objectionable various amounts of latency are for performers in live monitoring scenarios. Several popular instruments were used and the results of tests with wedge monitors are compared to those with in-ear monitors. It is shown that the audibility of latency is dependent on both the type of instrument and monitoring environment. This experiment shows that the acceptable amount of latency can range from 42ms to possibly less than 1.4ms under certain conditions. The differences in latency perception for each instrument are discussed. It is also shown that more latency is generally acceptable for wedge monitoring setups than for in-ear monitors.

1. INTRODUCTION

Over the past couple decades, digital audio gear has become increasingly common in nearly every area of audio production and reinforcement. Some concern has been raised that the latency inherent in a digital system can degrade the perceived quality of the audio and even affect the musician's ability to perform.

Some of the most common sources of latency are the use of digital buffers and filters. Even the process of sampling often includes a filter, and thus introduces some delay. As the number of filters is increased, the latency also increases. Therefore, it is expected that a large number of filter taps can easily create an audible delay, which may be perceived as comb filtering or even echo in the worst cases. Theoretically, even a 50μs delay (about 10 samples @192kHz) might be audible under certain conditions. If combined acoustically with the original sound, the resulting comb filter would have a null at approximately 10kHz. (Note: this delay would be equivalent to simply placing one speaker 1.7cm behind another speaker that is playing the same sound.) For an example, take one commercial audio processor that samples at 48kHz. It uses an ADC with a group delay of 37 samples and a DAC with a group delay of 29 samples, which results in a total latency of about 1.4ms. If this is combined

instantaneously with the original sound, it could result in comb filtering starting at frequencies as low as 357Hz. (Note: the severity of the comb filtering depends on the relative levels of the original sound and the processed sound at the ear.)

Some factors that might affect the perceived quality of a given amount of latency include the type of instrument played and the critical listening skills of the musician. Musicians on particular instruments, saxophone for example, might require more immediate feedback than others, such as keyboard, due to the physical coupling with the instrument. It may also be true that tonal instruments are more or less immune to perceived comb filtering than percussive instruments. As the latency is increased, the artifacts become less of a spectrum-altering phenomenon and more of a temporal perception issue. It is expected that some people are more adept at critical listening than others and will be able to hear small changes in the audio quality that others may have difficulty in perceiving until distinct temporal smearing is present.

This study is intended to quantify the amount of latency that is acceptable in a typical live-sound monitoring situation and to identify the variables that affect the variance of the results. This study is aimed at answering the following questions:

- *What are the different factors that affect latency perception in live monitoring?*
- *What are the differences in latency perception between two different monitoring situations: Wedge Monitors 4-6ft from the ear and In-Ear Monitors (IEM)?*
- *What are the differences in latency perception among different instrumentalists? Which musicians are more sensitive than others?*
- *Is there a difference between solo delayed monitoring and monitoring one's own delayed instrument while playing with a group of non-delayed musicians?*
- *How much latency can be present in a signal path before a musician will perceive an artifact in the audio signal?*
- *How much latency can be present in a signal path before a musician will perceive an actual delay in the signal?*

In order to answer these questions, there is an attempt to find the worst-case scenario and interpret the results with this in mind. In order to answer these questions with a high degree of confidence, this experiment uses 'critical' listeners rather than a general sample of the population. The subjects in this study are practicing musicians and care is taken to find professional musicians, where the term 'professional' is being defined as a person whose income is solely or partially supplemented by their musical ability.

2. METHODS

The amount of acceptable latency in a live sound application is interesting due to the implications on the design and use of professional audio products. Understanding what causes the perception of latency to change may help audio engineers determine what criteria to use when designing and employing products and systems. This experiment was designed with these objectives in mind.

2.1. Variables

The study was performed with the following variables:

- 19 Practicing Musicians, 11 of whom were professional musicians, determined to be critical listeners by initial boundary tests. (Note: the boundary tests were not factored into the final results.)
- 6 different instruments comprising a typical rock/pop band: Vocal, Saxophone, Electric Guitar, Keyboard with a piano patch, Electric Bass, and Drums
- Two different Monitoring Mechanisms: Wedges 4-6ft from the ear and In-Ear Monitors
- Two different Monitoring situations: solo monitoring and monitoring with a non-delayed metronome. This is meant to simulate playing with another musician whose signal has no latency

2.2. Tools

In order to evaluate a real-time system such as musician self-monitoring, a real-time switching mechanism needed to be used. It is important for the musician to have the ability to switch between samples quickly, easily, as many times as necessary, and with minimal artifacts. To meet these requirements an analog audio switchbox was designed and built for these experiments. The switchbox allows the user to select between 8 different channels with a very short cross-fade introduced during the switch.

In order to assess the subject's perception of latency, and to enable accurate comparison, the MUSHRA [1] methodology was used as a model. MUSHRA is a double blind multi-stimulus test for subjective quality measurements that involve medium to large signal impairments. This test provides a measure of the audio quality compared directly with a reference (the original, unprocessed audio signal). In a test involving intermediate impairments, subjects are asked to assess the basic audio quality of each sound. Since the subject can directly compare the effected signals, they can rate the effected signals with respect to each other, allowing

the experimenter to obtain grades with little random error. This permits a high degree of resolution in the assigned grades in the study.

This study uses the MUSHRA standard only as a model since the exact application is not perfectly applicable to the 'live' nature of the test. A live pseudo-MUSHRA test was developed and employed in this study. The differences are as follows:

- When grading a given sample, a known reference signal is not directly given as a means of comparison because, in a live situation, the instrument itself is the no-latency reference with which to compare. However, an analog reference sample is still hidden amongst the samples to assess the accuracy of the subject's responses
- The anchor is not a low-pass filtered version of the live signal since a low-pass filter is in no way related to the perception of latency. Instead, a sample that has a relatively large amount of latency is used as the anchor. The amount of latency that is introduced in the anchor sample is determined during an initial boundary test (described in Section 2.4).
- Instead of using a computer program that has sliders to rate the samples, the subject is given a paper grading sheet and is asked to mark an X where in the grading range the sample is perceived.

Additionally, this experiment is not double blind since the experiment proctor must manually set up the experiment for each trial. The official MUSHRA standard includes using a computer to randomly assign samples to the various channels. Although the test is not double blind, the proctor does not know what sample(s) the subject is currently testing and cannot give psychological cues to the subject; thus the single blind setup will not bias the results.

Since seven separate latencies in addition to the reference are needed, an eight channel digital mixer is used to introduce the latencies to the signal. It is important to note that when using digital equipment to introduce latency into a system that the propagation latency through the equipment must also be factored in. This particular device had a total throughput latency of 1.4ms. This latency is the smallest amount of delay that can be introduced to the signal; note that this corresponds to "0ms digital latency. "

2.3. Grading Process

In MUSHRA, the subjects score the stimuli according to a five-interval continuous quality scale. This scale consists of identical graphical scales that are long enough to give the subject enough quanta, typically 10cm or more, with an internal numerical description of 0 to 100. This scale is divided into five equal intervals with the following descriptors:

Excellent
Good
Fair
Poor
Bad

In a conventional MUSHRA test, the subject compares recorded audio samples that have varying severities of coding artifacts. However, the introduction of latency into a live signal has many more variables and potential sources of 'annoyance'.

The introduction of small latencies may not necessarily be perceived as a delay in the signal, although there are physical as well as psychophysical changes in the quality of the monitored signal. One physical effect is comb filtering within the monitored signal. One psychophysical effect may be the extra difficulty in judging when to initiate sound in the instrument by strumming, hitting, blowing, etc., based on the auditory feedback.

There are two primary phenomena that must be accounted for in the grading process: spectral artifacts arising from small amounts of latency, and the perception of audible delay. In the Pseudo Live-MUSHRA test, the MUSHRA grading standard is adapted to include these phenomena. Five similar descriptors are used, but their associated definitions are modified to include both phenomena:

Excellent: *Artifacts are imperceptible. Delay as well as artifacts cannot be identified.*
Good: *Some artifacts are perceptible, but not necessarily delay. The artifacts, though perceptible, are not annoying and do not contribute badly to musician's performance*.
Fair: *Delay and/or artifacts are perceptible. The delay and/or artifacts are slightly annoying, but in most cases would not affect musician's performance.*
Bad: *A considerable amount of delay is perceptible. The delay is annoying and is detrimental to musician performance.*
Horrible: *A musician can't work under these conditions*!

The range of the Excellent and Good categories generally covers the artifact phenomena; particularly comb filtering. The range of the Fair, Bad, and Horrible categories covers the delay phenomenon, which may be perceived as echo. Since the ordering and descriptions of the grading criteria are roughly proportional to the amount of latency to incite the phenomenon, the grading scheme is a fair one. In other words as you linearly increase the amount of latency in

the signal, the subject's responses follow the grading criteria linearly. This was verified during boundary tests.

Additionally, in an attempt to help the subject describe and grade the samples, more familiar descriptions were given in addition to the 'academic' descriptions:

- *The musician would buy a product that has the characteristics of the Excellent or Good category.*
- *The musician would use a product that has the characteristics of the Fair category if someone else owned the product or if it was installed in a venue. However, the musician would not voluntarily buy the product.*
- *The musician would not use a product that has the characteristics of the Bad or Horrible category, but would rather find an alternate product.*

The subjective scores are recorded in a numerical format. There are 5 levels of response and given the physical page in which to respond the levels of response get a score as follows:

Excellent	**100**
Good	**75**
Fair	**50**
Bad	**25**
Horrible	**0**

By physically measuring the position of a subject's mark, a numerical score between 0 and 100 is obtained.

2.4. Test Process

A few things were determined experimentally prior to the main study. These findings determined the hypothesis used to build the parameters of the study. The initial experimental design hypothesis and assumptions are as follows:

1. There is no one magnitude of latency that everyone will consider the threshold.
2. The subjects' responses for varying amounts of latency will follow a Gaussian distribution. In addition, sampling error will be the main cause of noise in the responses.
3. The range of latency magnitude in the eight samples needs to be determined on an individual and test-by-test basis. It is not possible to determine a perfect set of eight latencies that will allow every subject or even every instrument type to give responses in the range of the grading scale that also follows a Gaussian distribution.
4. The Gaussian distributions of a given instrument for different subjects will be somewhat similar; however the distributions for differing instruments will be very different.
5. A training phase is needed before recording test results to familiarize the subject with the differing types of artifacts and latency magnitudes. A training phase also helps reduce specific experiment related noise (error) in the results.
6. The subjects will be more sensitive to latency in the IEM mechanism rather than the floor monitor mechanism.
7. The subjects may be more sensitive to latency when playing with another musician whose sound is not delayed.
8. For subjects to be able to accurately rate the artifact portion of the grading scale, care needs to be taken to try to equally mix the original and monitored sound.

Due to numbers 1, 3, and 5 listed above, a boundary test method was designed to determine a range of latencies for each individual subject. A secondary purpose of this test was to train the subject. With the subject listening to a single channel, the proctor randomly chooses a delay setting and asks the subject to casually rate it. The subject responds given the same criteria as the actual experiment, only with a range of thumbs up to thumbs down for expediency. The angle of the thumb determines the approximate rating. The maximum latency is chosen as the minimum that consistently scores a thumbs down.

As well as determining the range of the latencies for the experiment, the boundary test also is used to generalize whether or not the musician is qualified as a 'critical' listener for this experiment. If a subject's responses are drastically different than his or her peers, or more deterministically if his or her threshold of perceiving actual latency is greater than the Precedence Effect, then the subject is likely not a critical listener.

The Precedence Effect, also known as the Haas effect, is a well-known interval of time whereby if two sounds are received within the interval, the brain considers them as one; outside the interval the brain detects two distinct sounds [2]. One major factor to note is the amplitude between the sounds. The stronger the delayed sound, the smaller the time interval. The most common values for this time interval are within the range of 25-50ms with the delayed sound up to the same amplitude. Most Precedence Effect experiments are based on speech and the subjects are a random sample of the population. It's possible that trained musicians could obtain lower values.

In this experiment, if the subject does not give a rating of less than or equal to Fair, with an introduced latency of 50ms, then the subject is considered an

outlier. This subject's results will not be included in the critical listening results.

2.4.1. Floor Wedges

There are a few constant conditions imposed on all subjects regardless of instrument. The first is that the distance between the floor wedge and the subject's ear is in the range of 4-6ft. This distance introduces a propagation delay of about 4.5-6.75ms at standard temperature.

Another imposed constant is the amplitude ratio of monitored and direct sound. In order to account for the most artifacts arising from small amounts of latency, the monitored and direct sound should be roughly equal in amplitude. However, having exactly equal amplitude from direct and monitored sound is not always a realistic situation for a musician on stage. The monitored sound is typically louder than the direct sound. Otherwise, the musician would be able to hear him or herself and the monitoring system would not be needed. Setting the sources equal only raises the Sound Pressure Level (SPL) by 3dB. This increase would not justify the use of monitoring for the gain in SPL. To account for this variable, the SPL is measured near the subject's ears both with and without the monitored sound. The increase in SPL that was comfortable for the subjects fell in the range of 4-5dB; higher than 3dB, but not high enough to eliminate the low latency artifacts such as comb filtering. Conversely the keyboard, electric guitar, and electric bass instruments have little or no direct sound. Although they will have no physical artifacts such as comb filtering, psychophysical effects may still be present.

In the floor wedge trials, the eight latencies included 0ms (analog) and seven other linearly increasing, digitally delayed latencies.

2.4.2. In-Ear Monitors

With an increasing number of musicians using In-Ear Monitors (IEM) on stage, it is important to both quantify the latency requirements for IEM and compare the results to that of the floor wedges. With the monitor source now directly in the subjects' ear, propagation delay is practically eliminated. However, the subject now directly controls the monitoring level. Subjects were instructed to set the volume to a comfortable monitoring level. Fortunately, since the same musicians are conducting the IEM experiment as the floor wedges experiment, it is reasonable to assume that their levels for the IEM experiment will be similar to the floor monitor version and that the level is more or less representative of on-stage parameters.

There are more issues to be aware of in regard to artifacts in the IEM experiment. Particularly with instruments that have physical coupling with the body such as vocalists or saxophonists, there is likely an increased ability to distinguish artifacts from combined versions of direct and delayed sound. Additionally, there are artifacts arising from the occlusion effect, and the psychological discomfort of the delayed sound not time aligning with the vibrations of the instrument.

In the IEM trials, the eight latencies used included 0ms analog, 1.4ms sampling latency (0ms digital) and six other digitally delayed linearly increasing latencies starting from 1.4ms.

2.4.3. Solo Monitoring

The first experiment is the least complicated; the musician simply plays music of his/her choice while first monitoring through floor wedges and then through the IEM mechanism. The delays determined in the boundary test are randomly applied to the 8 channels of the switchbox. The musician rates the channels accordingly.

2.4.4. Delayed Self Monitoring with Non-Delayed Metronome

To determine the sensitivity to latency while playing along to a non-delayed source, an experiment including a constant time source is included. Adding another live musician would add a significant variable to the experiment and is not practical for data analysis. To simplify the idea but maintain the utility, an experiment is conducted with a delayed instrument signal and a non-delayed metronome. The metronome speed is held at a constant 120 BPM for all subjects and all experiments. The subject adjusts the sound pressure level of the metronome to a comfortable level.

3. ANALYSIS

By adjusting the sample latency range to the specific subject, there are many different magnitudes of latency being tested in the breadth of the experiment. The latency values range from 0ms (analog) to 70ms in various linear permutations of 8 values. One should note that the IEM experiment does not have 8 linearly related latencies, but rather 7, since there are samples of both the analog signal and the "0ms digital" latency.

Since each subject has two different sets of latency values and there are 19 musicians, there is a potential for 38 completely different sets of latency. In order to correctly compare results, the overall implied trends of every subject must be noted.

From both the boundary tests and the output results, it can be shown that each subject's noisy output can be estimated by a cumulative Gaussian function.

Some notable examples of subject's raw data following a relative noisy cumulative Gaussian trend are shown in Figs. 1-4.

Although there are phenomena that cause some of the data points to fall outside of a cumulative Gaussian function, a general trend can be seen by observing the data. This trend allows the data to be directly compared to each other without having explicitly identical stimuli.

The overall trend can be seen when including all responses on one graph as shown in Fig. 5.

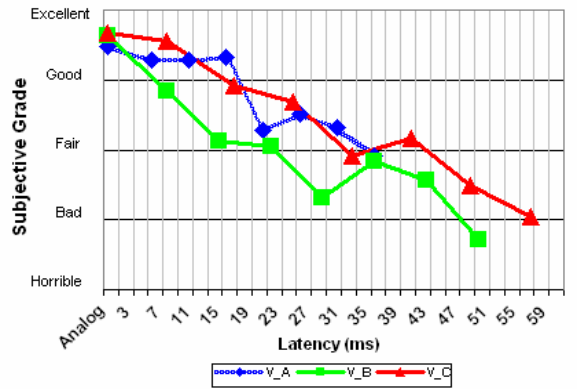


Figure 1: Vocals with Metronome Wedge Monitor Raw Scores

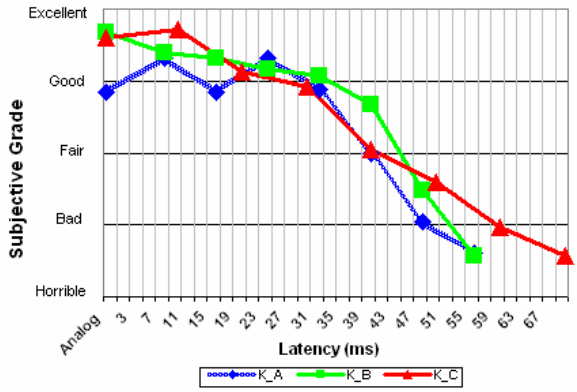


Figure 2: Keyboard with Metronome Wedge Monitor Raw Scores

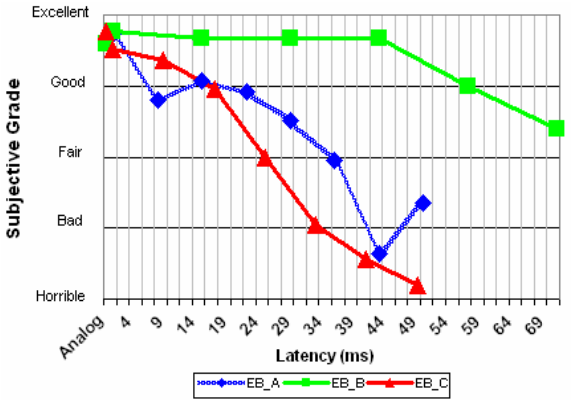


Figure 3: Electric Bass In Ear Monitor Raw Scores

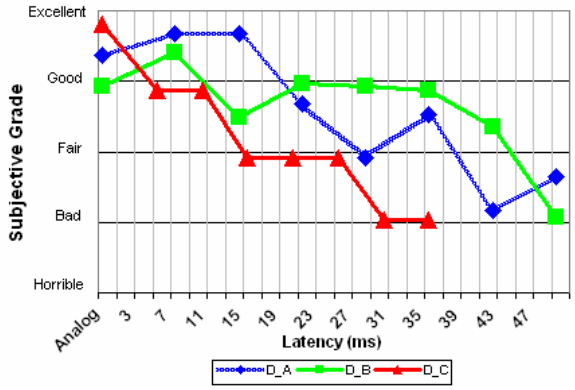


Figure 4: Drums with Metronome Wedge Monitor Raw Scores

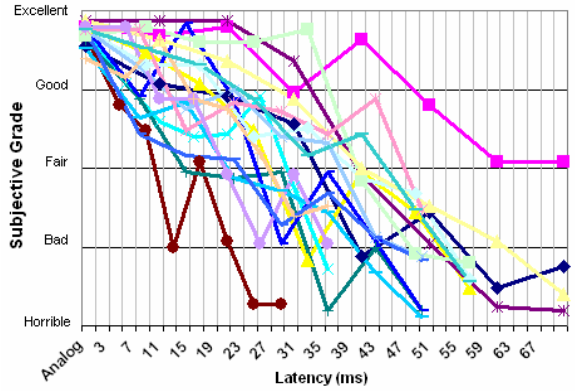


Figure 5: All Wedge Raw Scores

NOTE: The outliers are included in all of the above graphs.

3.1. The Implication of Different Criterion

A simple glance at Figs. 1-4 reveals two things:

- Instrumentalists' criterion to judge their perception of latency vary from instrument to instrument
- Within a specific instrument, the criterion is generally very similar amongst various musicians.

In other words, sensitivity to latency is more strongly dependent on the instrument rather than the individual subject. To support this claim, consider the following: in this study, there are two musicians that performed on two separate instruments. The ratings that the subjects gave were consistent with the instrument type rather than consistent with their own ratings on the other instrument.

Comparison of the characteristics of SUBJECT C's performance on Electric Bass and on Keyboard (Note- this is the same person):

- The trend curve follows the respective instrument's overall trend.
- The thresholds of Good and Fair are consistent with the instrument's overall trend.
- The critical thresholds of Good and Fair for Electric Bass are about 18 and 28ms (Fig. 6) respectively whereas the critical threshold of Good and Fair for Keyboard is about 30 and 43ms respectively (Fig. 9). This exhibits an average criterion increase of 60% from Electric Bass to Keyboard.

Note: These example values were taken from the Wedge Monitor without Metronome graph. The definition of critical threshold is the value in ms at which the curve first crosses the given rating with respect to the curve trend. For example, given the case of Drums IEM subject C (Fig. 13), the positive trend in the 0-6ms range would be ignored and the Good critical threshold would be about 9ms even though 23ms also crosses the Good range.

Comparison of the characteristics of SUBJECT A's performance on Keyboard and SUBJECT B's performance on Drums (Note- this is the same person):

- The trend curve follows the respective instrument's overall trend.
- The thresholds of Good and Fair are consistent with the instrument's overall trend
- The critical threshold of Good and Fair for Drums is about 12 and 48ms respectively (Fig. 12) whereas the critical threshold of Good and Fair for Keyboard is about 21 and 41ms (Fig. 9). There is an increase in Good rating of about 75%, but there is a reduction of Fair rating by about 15%.

- Note: These example values were taken from the Wedge Monitor without Metronome graph.

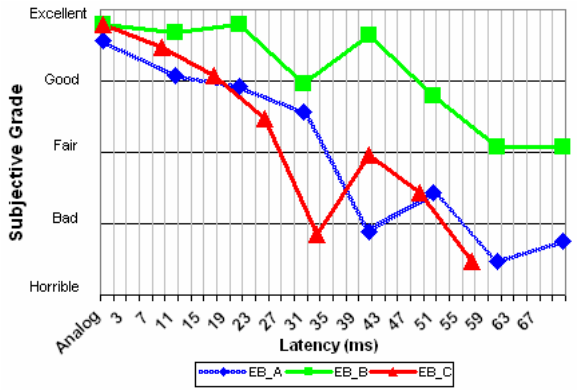


Figure 6: Electric Bass Wedge Monitor Raw Scores

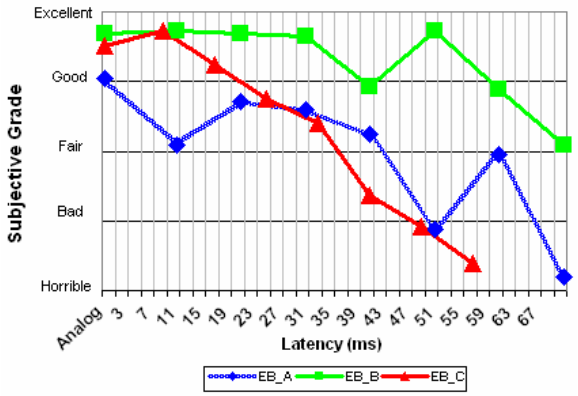


Figure 7: Electric Bass with Metronome Wedge Monitor Raw Scores

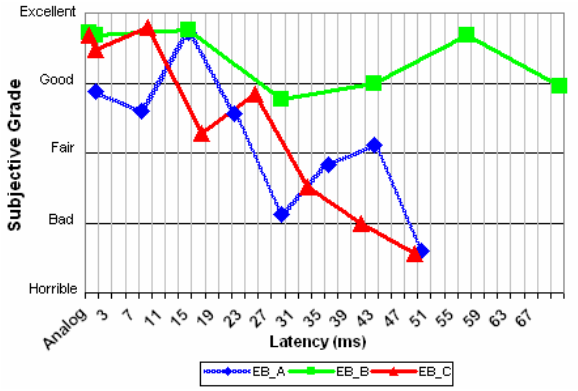


Figure 8: Electric Bass with Metronome IEM Raw Scores

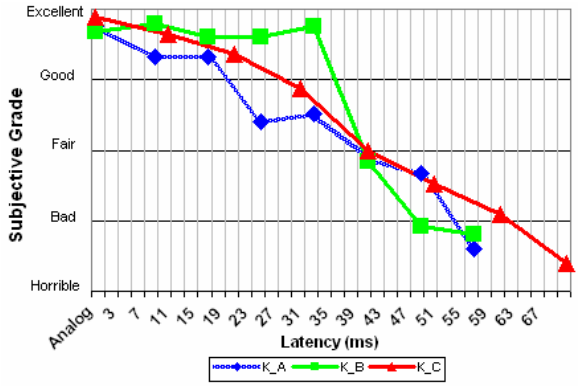


Figure 9: Keyboard Wedge Monitor Raw Scores

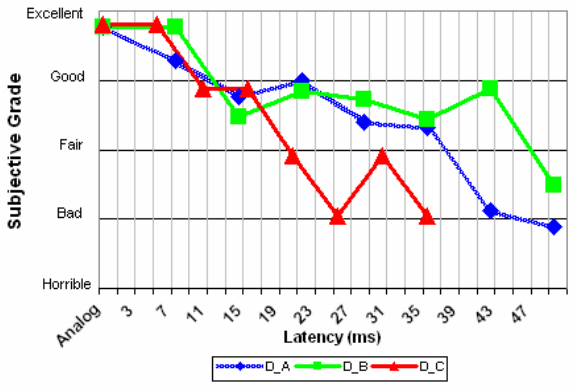


Figure 12: Drums Wedge Monitor Raw Scores

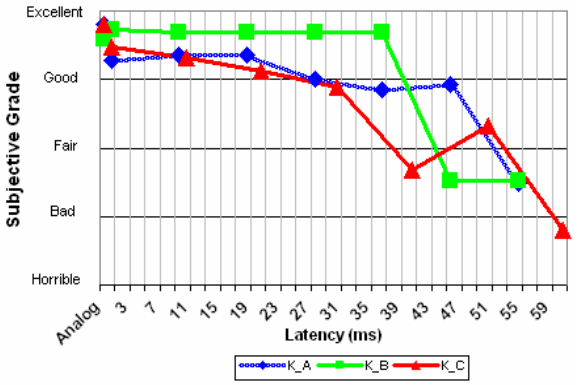


Figure 10: Keyboard IEM Raw Scores

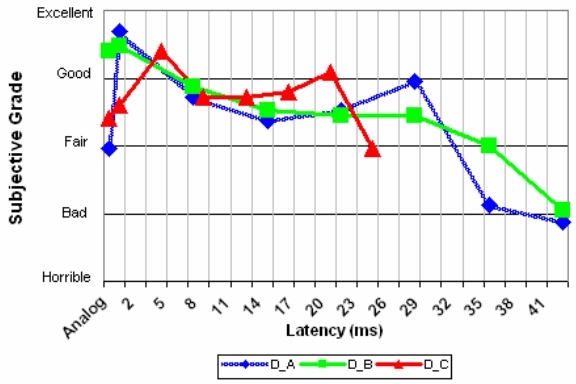


Figure 13: Drums IEM Raw Scores

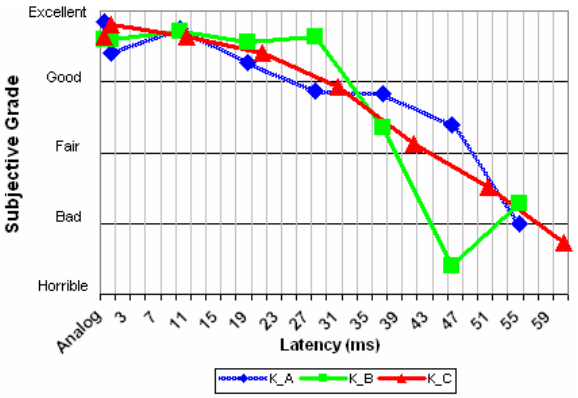


Figure 11: Keyboard with Metronome IEM Raw Scores

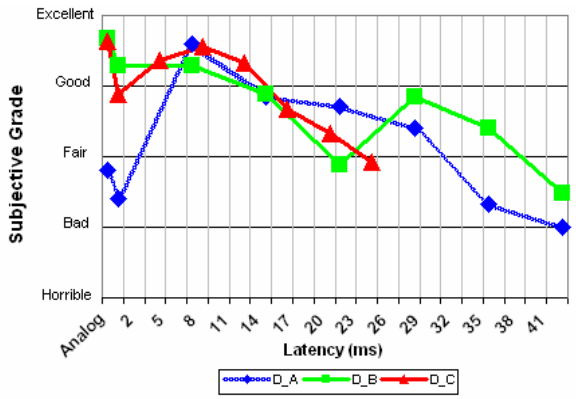


Figure 14: Drums with Metronome IEM Raw Scores

When analyzing Figs. 6-14 as well as the Figs. 15-26 presented in the following section, there is an interesting phenomenon to note. There is an implication that while in the Wedge Monitor mechanism no latency is the highest rated sample, in the IEM mechanism the highest rated sample can move from no latency to a higher latency. This is suggested considering only the samples containing no metronome so as to not include additional independent variables.

Overall 9 out of 17 subjects (52.9%) reported that the higher rated sample moved from no latency to some latency. Although this is not a very high percentage, consider that 0 subjects (0%) followed the opposite trend, reporting that the highest rated sample moved from some latency to no latency.

It could be suggested that when an instrumentalist monitors performance with IEM, a latency that is roughly equal to that of the distance in free air of his or her ear to the instrument may be preferable.

3.2. The Impact of a Non-Delayed Metronome

When analyzing Figs. 6-14 as well as the Figs. 15-26 presented below, notice that one distinguishable difference between the graphs with and without a metronome is the position of the highest rated latency. In many cases, the highest rated latency is no longer 0 ms but has shifted up by an amount. In other cases, the general shape of the curve has simply shifted to the right, indicating that more latency is necessary to incite the same subjective rating with a metronome than without. Of the 35 responses, the following is observed:

- 18 responses (51.4%) exhibit a shift in the highest rated latency from 0ms to different low latency
- 4 responses (11.4%) exhibit a shift to the right in the general shape of the curve in the lower latency area (<15ms).
- 13 responses (37.1%) show no low latency (<15ms) change.

Overall, 62.9% of trials exhibited a trend shift in the lower latency range. This implies that when playing with a non-delayed metronome or with another non-delayed musician, perhaps some latency is preferable to none.

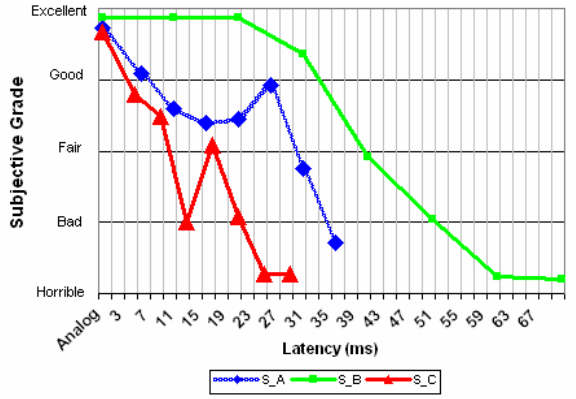


Figure 15: Saxophone Wedge Raw Scores

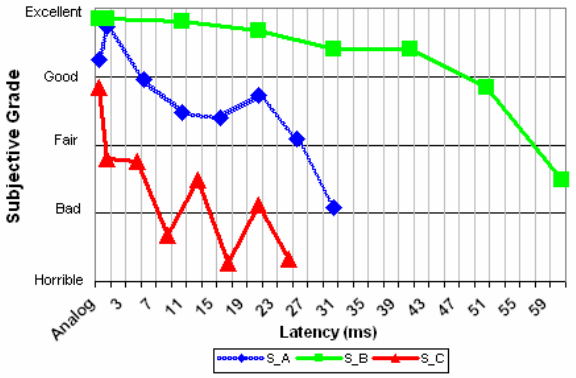


Figure 16: Saxophone IEM Raw Scores

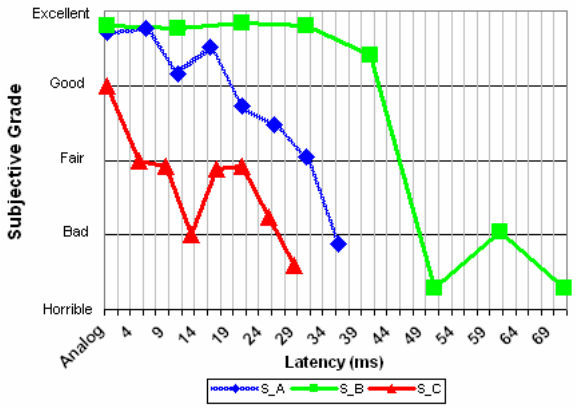


Figure 17: Saxophone with Metronome Wedge Monitor Raw Scores

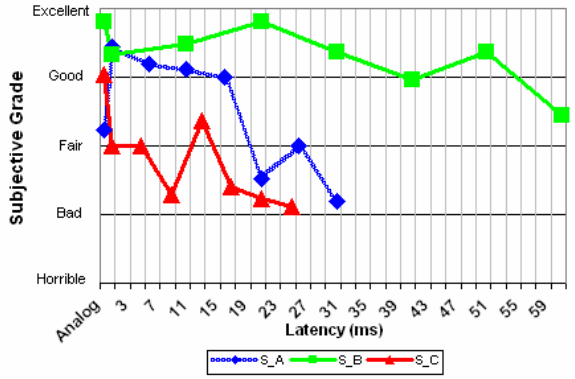


Figure 18: Saxophone with Metronome IEM Raw Scores

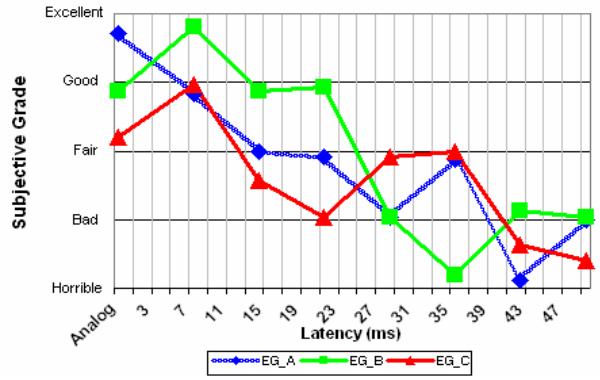


Figure 21: Electric Guitar with Metronome Wedge Monitor Raw Scores

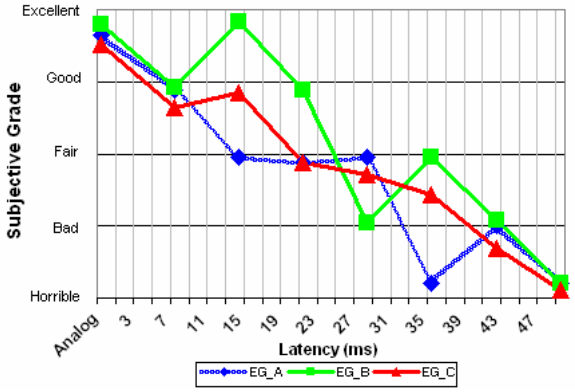


Figure 19: Electric Guitar Wedge Monitor Raw Scores

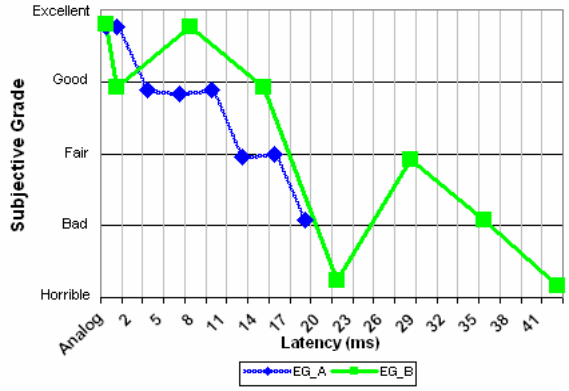


Figure 22: Electric Guitar with Metronome IEM Raw Scores

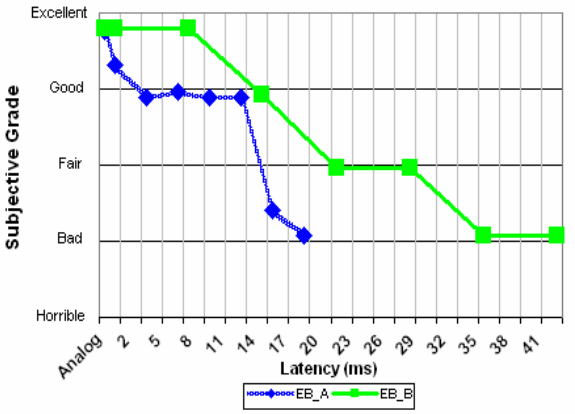


Figure 20: Electric Guitar IEM Raw Scores

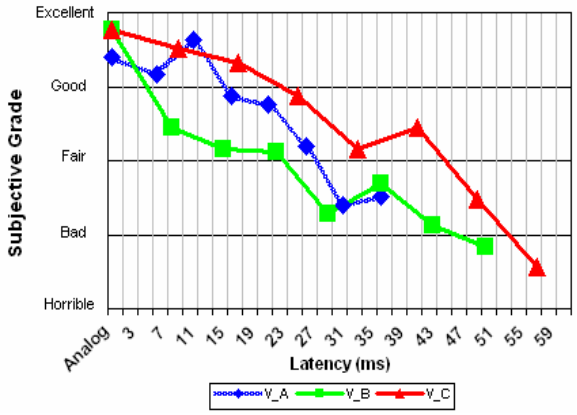


Figure 23: Vocals Raw Wedge Monitor Scores

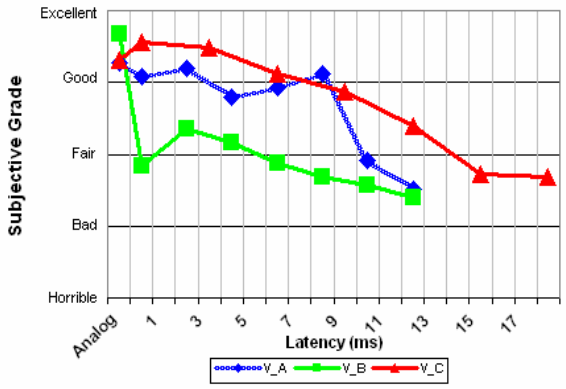


Figure 24: Vocals IEM Raw Scores

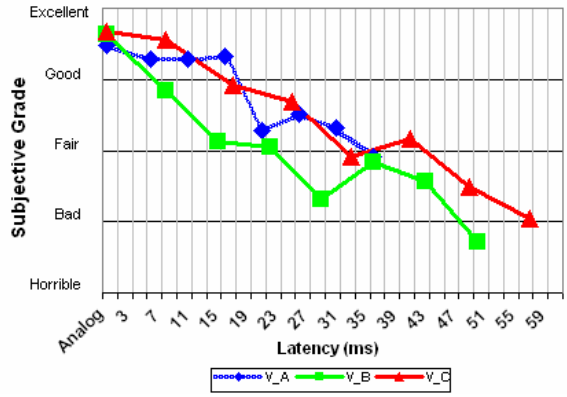


Figure 25: Vocals with Metronome Wedge Monitor Raw Scores

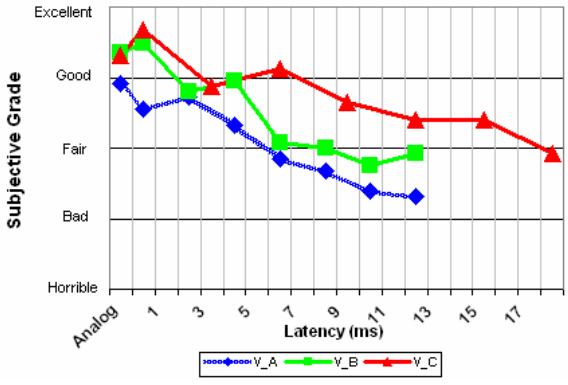


Figure 26: Vocals with Metronome IEM Raw Scores

3.3. Data Trend Analysis

It has been shown that there is a trend in the data, which allows it to be analyzed and compared even though the latency amounts in each trial are not identical. Furthermore, the data points should be preserved rather than estimated to exploit any additional unaccounted for phenomenon. However, it has also been shown that there is different latency criterion correlated to which instrument is under test.

Linearly interpolating scores between the actual scores safely augments the data for analysis. All effort is made to use the actual data in analysis rather than the interpolated data where possible. The interpolation includes all real integer latencies necessary to cover subjective grade range of interest, 50 (Fair) – 100 (Excellent).

3.3.1. Subjective Cross Sections

In order to evaluate the quality of various amounts of latency, the latency thresholds of Good and Fair are of interest. The critical threshold described in Section 3.1 is used to define this latency. The latency is estimated within 0.25ms based on the interpolated data for the four configurations: wedge without metronome, wedge with metronome, IEM without metronome, and IEM with metronome.

Data that has similar variables and trends can be combined. In order to most accurately analyze the data, all responses from each monitoring mechanism per instrument are combined. For example, both the data with and without metronome for keyboard in the wedge monitor mechanism are used to estimate the thresholds for Wedge Monitors; the Wedge and IEM mechanisms are not combined nor are any separate instruments.

When combining this data, the outliers (subject B for both electric bass and saxophone) were not included. Not only are these subjects identified as outliers by the definition of a critical listener outlined in Section 2.4, but their trends also do not match the other respective instrumentalists.

3.3.2. Lower Bounds using Confidence Intervals

This experiment is aimed at identifying and comparing thresholds of latency in live sound monitoring; more specifically it is aimed at finding a lower bound, or worst case, trend of these values. To get an estimated lower bound trend, confidence intervals are used. The confidence interval, calculated from the given set of sample data, gives an estimated range of values that is likely to include the mean of any given set of random listeners. Since the subjects in this study are not truly random but rather critical listeners,

the confidence interval estimates the range not for given set of random listeners but for a given set of *critical* listeners. This is one way that the experiment considers the estimate a lower bound.

The confidence interval gives a range of values both above and below the sample mean of the latency scores. The width of this interval determines how precise and correlated the data is. However, since only the lower bound latency is of interest, the whole interval does not need to be calculated. The lower bound of the confidence interval is determined and plotted at all confidence levels 0-100%. The transition width, the range of latencies it takes to go from a near 100% level to a near 0% level, implies the precision and correlation of the data used to plot the given confidence curve. The more narrow the curve, the more precise and correlated the data is. This curve allows the reader to decide, based on his or her own criteria, the thresholds of Good and Fair given the confidence level and the transition width of curve. See Section 4.1 for the figures.

To determine the difference between the wedge monitor and IEM mechanisms, a horizontal cross section at various confidence levels is plotted for all instruments. This allows the reader to identify trends in the acceptable amounts of latency in all mechanisms. See Section 4.2 for the figures.

3.3.3. Separating Trends from the Actual Numbers

In addition to the analysis of Section 3.3.2, there is merit in trying to extract individual instrument trends from the actual latency values. (It is important to note that the following methods would be more accurate with an increased number of subjects. The trends and the methods used to obtain them are simply suggestions and are not confidently portrayed as fact.) To identify the trends, one variable is held constant while the other is analyzed. One analysis includes viewing the change from a Good to a Fair rating with a constant of listening mechanism; the other includes viewing the differences in IEM and Wedge mechanism with a constant subjective rating.

A visual representation of the increase in latency required for a change of a Good Rating to a Fair Rating (Δt_{G2F}) can be obtained by finding percent increases in latency for each instrument with respect to monitoring mechanism (MM). This allows the reader to easily see the change in sensitivity with respect to Subjective Rating among instruments, in other words how much more latency would need to be added to a system before the musician will change their opinion on the quality.

$$\Delta t_{G2F} = \frac{MM_{Fair} - MM_{Good}}{MM_{Good}} \quad (1)$$

Similarly, a visual representation of the difference in latency amounts for each monitoring mechanism (Δt_{I2W}), IEM to wedge, can be obtained by finding percent change in latency for each instrument with respect to subjective rating (SR). This graph allows the reader to easily see the change in sensitivity with respect to monitoring mechanism, in other words how likely a musician is to change the opinion on the latency in the system based on an environmental change such as change in monitoring mechanism.

$$\Delta t_{I2W} = \frac{SR_{Wedge} - SR_{IEM}}{SR_{IEM}} \quad (2)$$

The combination of the two representations isolate the two forms of latency sensitivity for each instrument: sensitivity to absolute latency amounts, which implies independent criteria, and sensitivity to change in monitoring mechanism, which implies dependent criteria.

Both sensitivities have been separated from the actual latency threshold magnitudes. This means that it is possible for an instrumentalist to have a high subjective threshold. In other words the subject can accept a large amount of latency before meeting subjective criteria, but at the same time have high sensitivity to criteria and environmental change. In accordance to the principles of detection theory, if the criteria can be extracted from the actual data, the sensitivity can be described independent of personal bias or personal subjective criteria [3]. In this case, there are three measures of 'bias': bias due to instrument configuration, bias to environmental (monitoring) configuration, and bias due to the individual subjects giving either liberal or conservative subjective ratings. Subject bias cannot be separated at this point in analysis since the individual data has already been combined; however, we can attempt to separate the other biases by separating the trends from the data and viewing the differences objectively.

The following describes a proposed method to find the 'General Sensitivity'. First, the absolute value of the difference in Good and Fair percent increase for 'IEM to Wedge' mechanism is obtained using Equation 1. This represents sensitivity of subjective rating with respect to environmental change.

$$\Delta Eq_1 = \left|\Delta t_{G2F}(IEM) - \Delta t_{G2F}(\text{Wedge})\right| \quad (3)$$

where ΔEq_1= the percent difference of IEM and Wedge mechanisms for a Good to Fair subjective increase
Δt_{G2F}(IEM) = Equation 1 evaluated for IEMs
Δt_{G2F}(Wedge) = Equation 1 evaluated for Wedges

Next, the absolute value of the IEM to Wedge percent change difference for 'Good to Fair' is obtained using Equation 2. This represents sensitivity of environmental change with respect to subjective rating.

$$\Delta Eq_2 = \left|\Delta t_{12W}(Good) - \Delta t_{12W}(\text{Fair})\right| \quad (4)$$

where ΔEq_2 = the percent difference of Good and Fair subjective increase for a IEM to Wedge change
$\Delta t_{l2W}(Good)$ = Equation 2 evaluated for Good
$\Delta t_{l2W}(Fair)$ = Equation 2 evaluated for Fair

Both sensitivities need to be accounted for in the general sensitivity calculation, thus the un-weighted average (Λ) is taken.

$$\Lambda = \frac{\Delta Eq_1 + \Delta Eq_2}{2} \quad (5)$$

To describe the general sensitivity with respect to instrument type, the instruments need to be compared to each other. The mean of the individual un-weighted averages calculated by Equation 5 for each instrument is taken to provide a normalization factor (NF) with which to use as a baseline to graph the general sensitivity amongst instruments.

$$NF = \frac{\sum_{i=1}^{n} \Lambda_i}{n} \quad (6)$$

where N = Number of instruments
Λ_i = Individual instrument average sensitivity from Equation 5

Finally, the average for each given instrument, calculated by Equation 5, is divided by the normalization factor (NF) from Equation 6 to obtain the general sensitivity measure.

$$\text{General Sensitivity} = \frac{\Lambda}{NF} \quad (7)$$

The General Sensitivity has no units. General Sensitivity allows the instruments to be compared to each other. Additionally, it is used as a measure of confidence in describing absolute thresholds. The general sensitivity measure attempts to extrapolate when a musician playing a particular instrument type is likely to change his or her opinion if independent variables that cannot be accounted for in this experiment are changed. A higher general sensitivity signifies that the instrumentalist is more likely to change his or her opinion of latency when a criterion such as listening environment changes. Conversely, a lower general sensitivity signifies that the instrumentalist is less likely to change his or her thresholds of latency when criterion changes.

4. RESULTS

4.1. Individual Instrument Confidence Curves, All Configurations

The following graphs plot the confidence curves for each individual instrument.

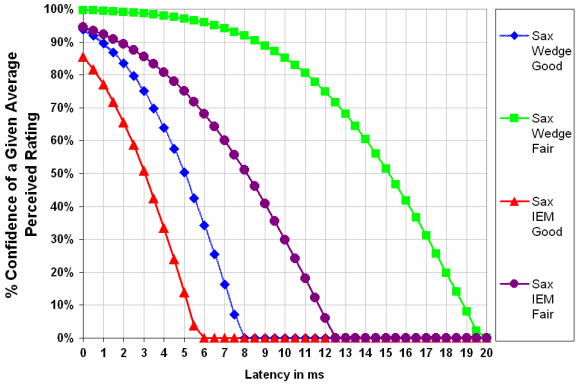


Figure 26: Saxophone Confidence Curve, all Configurations

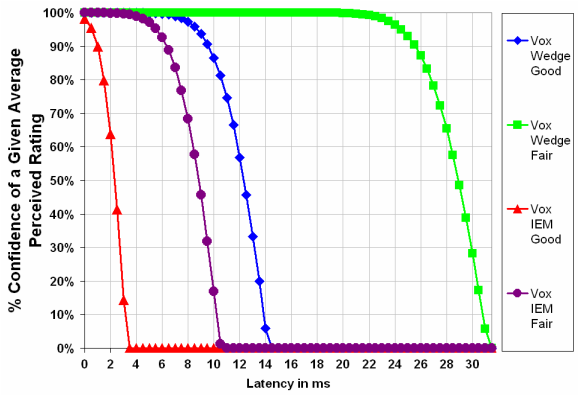


Figure 27: Vocal Confidence Curve, All Configurations

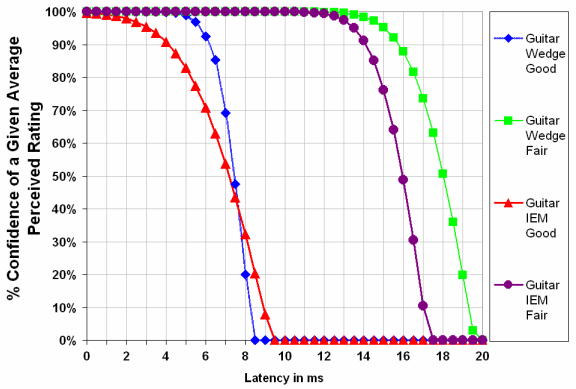


Figure 28: Electric Guitar Confidence Curve, All Configurations

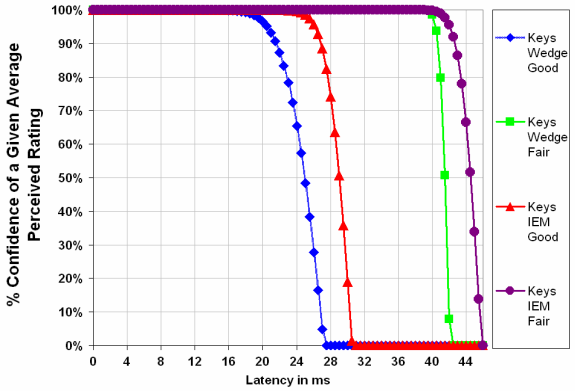


Figure 31: Keyboard Confidence Curve, All Configurations

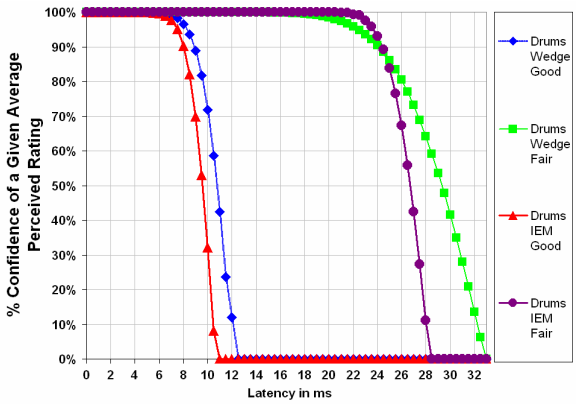


Figure 29: Drums Confidence Curve, All Configurations

Figs. 26-31 demonstrate that the Good rating for both mechanisms has a higher threshold than the Fair rating for both mechanisms except for the vocals. This suggests that vocalists may be more sensitive to the monitoring mechanism than other musicians.

More saxophone musicians would need to be tested to get a more accurate result. The transitions width of the confidence levels is too wide, implying that there is too much variance in the results.

4.2. Individual Configurations, All Instrument Confidence Curves

The following graphs (Figs. 32-35) plot the confidence curves for each subjective rating and monitoring mechanism combination for all instruments.

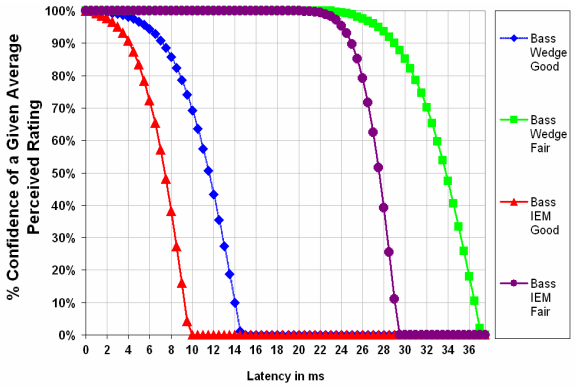


Figure 30: Electric Bass Confidence Curve, All Configurations

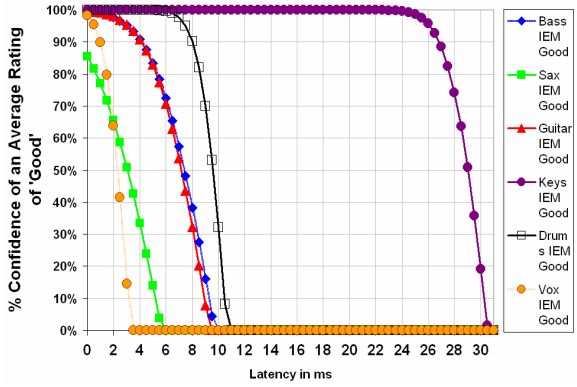


Figure 32: IEM Good Rating, All Instruments

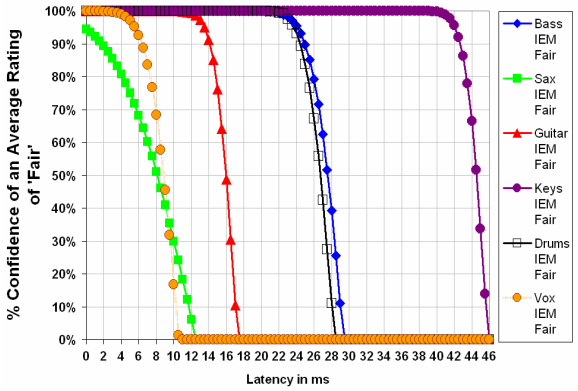


Figure 33: IEM Fair Rating, All Instruments

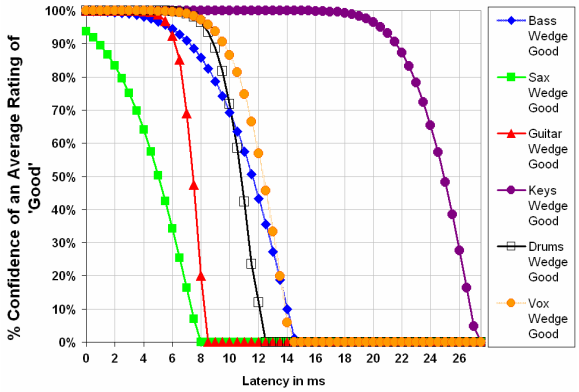


Figure 34: Wedge Good Rating, All Instruments

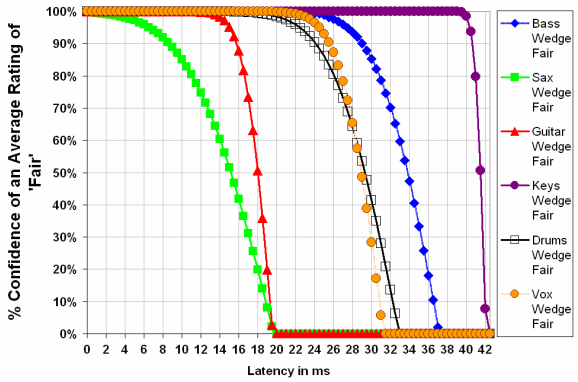


Figure 35: Wedge Fair Rating, All Instruments

The keyboard subjective threshold for all mechanisms and all ratings is significantly higher than the thresholds of the other instruments. The keyboard is the least critical of latency when giving subjective grade thresholds.

4.3. Comparison of Instruments at a Given Confidence Level

Fig. 36 shows the 85% confidence level cross-section of the graphs in Section 4.2 to more easily compare critical thresholds for a given subjective rating. The 85% confidence level is used to show better detail for the saxophone results, since the variance in the saxophone data is high.

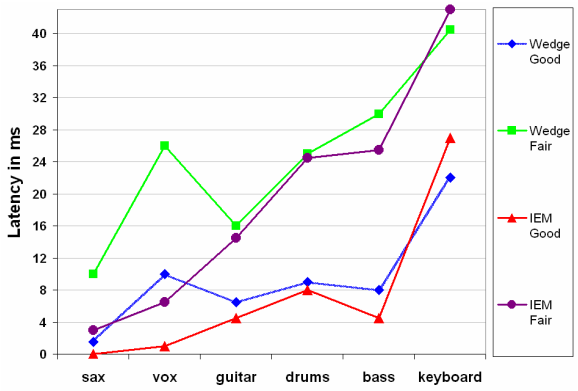


Figure 36: Instrument Comparisons at 85% Confidence Level

Latency (ms)	Sax	Vocals	Guitar	Drums	Bass	Keys
IEM Good	0	1	4.5	8	4.5	27
Wedge Good	1.5	10	6.5	9	8	22
IEM Fair	3	6.5	14.5	24.5	25.5	43
Wedge Fair	10	26	16	25	30	40.5

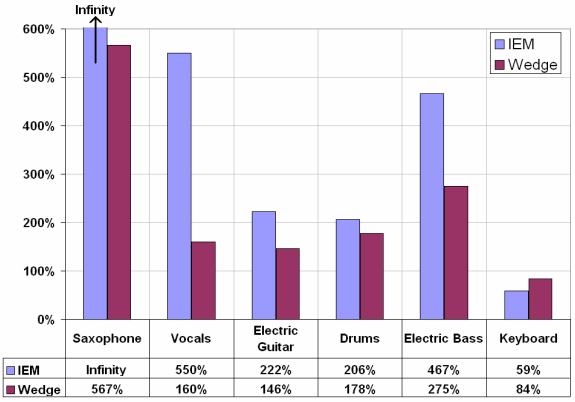
Table 1: Instrument Comparison Table at 85% Confidence Level for Clarity

Although Fig. 36 shows that the saxophone has a more critical subjective rating threshold than all other instruments, due to the variance in the data, more musicians would need to be tested to completely confirm this claim.

These graphs make it apparent that the vocalists follow a trend similar to other instrumentalists for the IEM mechanism, but have an increase in critical latency threshold for the Wedge Monitor mechanism that is not similar to other instruments. Another interesting point is that unlike all other instruments, the keyboardist can accept more latency in the IEM than the Wedge mechanism.

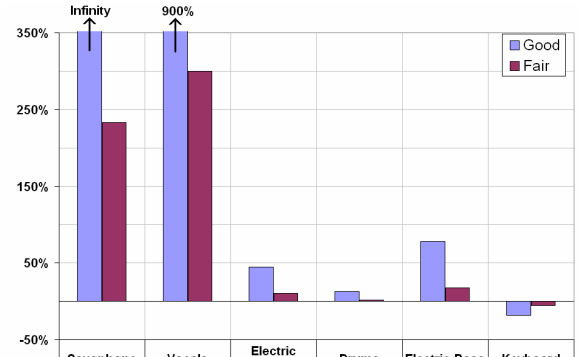
4.4. Instrument Sensitivity Measures

The following figures describe the sensitivity of each instrument to independent variable change. (Note: a division by 0ms causes a result of infinity)



	Saxophone	Vocals	Electric Guitar	Drums	Electric Bass	Keyboard
IEM	Infinity	550%	222%	206%	467%	59%
Wedge	567%	160%	146%	178%	275%	84%

Figure 37: Good to Fair Latency Threshold Percent Increase at 85% Confidence Level

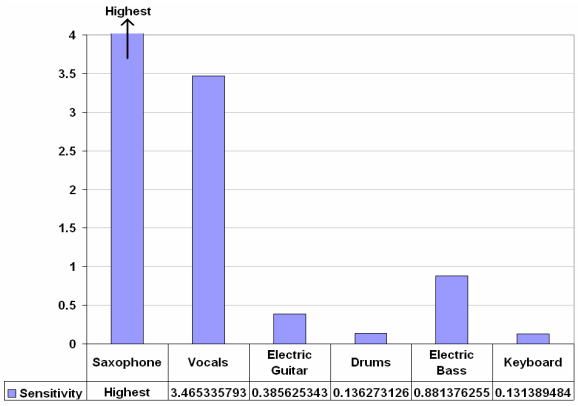


	Saxophone	Vocals	Electric Guitar	Drums	Electric Bass	Keyboard
Good	Infinity	900%	44%	13%	78%	-19%
Fair	233%	300%	10%	2%	18%	-6%

Figure 38: IEM to Wedge Latency Threshold Percent Increase at 85% Confidence Level

The high saxophone general sensitivity is due to the number infinity being introduced into the calculation due to division by zero. It is not necessarily explicitly implied that the saxophone is highly sensitive to changes in perception due to criteria or environmental changes.

Vocalists have a high general sensitivity and are thus more likely to change their perception of latency based on variable change. This means that the thresholds described in Section 4.3 are not precise in a wide variety of situations.



	Saxophone	Vocals	Electric Guitar	Drums	Electric Bass	Keyboard
Sensitivity	Highest	3.465335793	0.385625343	0.136273126	0.881376255	0.131389484

Figure 39: General Latency Sensitivity for All Instruments

Drummers and keyboardists have a low general sensitivity and are thus likely to perceive latency very similarly in a wide variety of situations.

5. DISCUSSION

Sensitivity to latency is more strongly dependent on instrument rather than the individual subject.

This is suggested with the single performer, multiple instrument isolated case discussed in Section 3.1. Additionally, this claim is strongly supported in all of the figures in this paper.

It could be suggested that when an instrumentalist monitors performance with an IEM, a latency that is roughly equal to the distance between his or her ear and the instrument may be preferable.

This claim can only be suggested rather than concluded since this particular test paradigm is not set up to measure the smaller latency quanta necessary to get an accurate preference rating. A more critical test method, such as that recommended in ITU BS.1116, would need to be used to verify this claim [4].

When playing with a non-delayed metronome or with another non-delayed musician, some latency may be preferable to none.

As shown in Section 3.2, 62.9% of trials exhibit a trend shift to a higher subjective threshold.

Consider also a study by Chris Chafe and Michael Gurevich on Network Time Delay and Ensemble Accuracy [5]. The study suggests that the sweet spot for musicians to play with each other in tempo is 11.5ms. This, of course, corresponds to a distance between the musicians in free air. Statistically, a delay of less than 11.5ms caused 74% of musicians to speed up, and a delay of more than 11.5ms caused 85% of the musicians to slow down. Additionally, they determined that the longest propagation delay in free air for a typical ensemble is around 20ms.

Vocalists are much more sensitive to latency in the IEM mechanism than the Wedge monitor mechanism.

As can be seen in Fig. 27, vocalists tend to have a lower latency threshold for a 'Fair' IEM mechanism than a 'Good' Wedge monitor system. This is in contrast to the results from other instruments (Figs. 26 and 28-31).One possible explanation for this is when using a Wedge monitor, there is natural reverb added from the room. When reverb is added, latency tends to blend itself in with the reverb. Moreover, it can be suggested that many vocalists prefer to hear their voice with audio processing effects such as reverb. This means that the latency with room reverb is actually a desired effect. When the vocalist uses an IEM, a different expectation results. The vocalist does not anticipate hearing reverb since the room sound is not longer present. This directly affects the vocalists' criteria of subjective latency grading. Perhaps this difference between vocalists and other instrumentalists due to reverb is the primary variable that makes the vocalists general sensitivity to latency higher than other instruments.

More saxophone musicians should be tested to get a more accurate result. The transition width of the confidence levels is too wide implying that there is too much variance in the experiment results.

Since one saxophone subject was deemed an outlier and not used, out of two subjects one subject was highly critical of latency and one subject was moderately critical of latency. This results in high data variance. The mean and standard deviation sets for saxophone data are as follows:

- Wedge, Good: Mean = 7.9 σ = 8.5
- Wedge, Fair: Mean = 19.7 σ = 13.4
- IEM, Good: Mean = 5.7 σ = 7.8
- IEM, Fair: Mean = 12.5 σ = 13

When taking confidence intervals with these numbers, some confidence levels returned negative values. This implies that even at 0ms latency, the result is not 100% confident. More subjects are needed to get a more accurate idea of a typical saxophonists' performance. However, since this study is aimed at finding the critical latencies, it could be concluded that the results of the highly critical subject can roughly represent the worst case for the typical saxophonist. On the other hand, this highly critical listener could be an outlier with respect to even critical listeners. This musician's thresholds are offered with the advice of interpreting these numbers with caution.

Critical Saxophonist's Thresholds:

- Without metronome:
  - Wedge, Good: 4.4ms
  - Wedge, Fair: 11.4ms
  - IEM, Good: 0ms
  - IEM, Fair: 1.4ms
- With metronome:
  - Wedge, Good: 0ms
  - Wedge, Fair: 5.4ms
  - IEM, Good: 0ms
  - IEM, Fair: 1.4ms
- Average Thresholds
  - Wedge, Good: 2.2ms
  - Wedge, Fair: 8.4ms
  - IEM, Good: 0ms
  - IEM, Fair: 1.4ms

The keyboard is the least critical of latency when giving subjective grade thresholds. The keyboard subjective threshold for all mechanisms and all ratings is significantly below the thresholds of the other instruments.

The keyboard is clearly the least critical of latency magnitude; this result is expected. A musician who plays keyboard is accustomed to playing with a variety of synthesizer patches. Many patches have a very slow attack time. These patches feel like there is a lot of delay between the finger attack and the resulting sound. Keyboard musicians have adapted to this latency and can easily play in time by slightly anticipating the actual attack time.

With the exception of the keyboard, the instruments in the IEM mechanism require a larger latency increase to change subjective criteria thresholds than in the Wedge mechanism.

This could simply be due to the fact that since the latency required for a good rating in IEM is lower, the ratio ends up being higher due to a consequence of calculation. Indeed, the criteria for a Good rating in IEM will be more stringent; the criteria for Good means the subject can hear artifacts but not delay. Since the monitored source is now in the subject's ear, artifacts arising from phenomena such as comb filtering will be more prominent with lower amounts of latency.

The keyboard is less sensitive to latency in the wedge mechanism. It takes a larger, but similar, increase in latency to change subjective criteria thresholds (from Good to Fair) in the Wedge mechanism than it does in the IEM mechanism for the Keyboard.

The keyboard is an instrument that provides no direct sound. Therefore, there are no artifacts due to lower amounts of latency since there is no direct sound. It is possible that keyboardists are using the additional grading criteria more than the primary criteria given in Section 2.3. If so, this may explain why the increase from Good to Fair is greater than that of the other instruments.

The drummers do not show a very large criteria change from IEM to Wedge, meaning the drummers are not very sensitive to configuration changes.

The amount of latency seems to be more important to a drummer when providing threshold of quality than the way in which the latency is added to the signal. This may be because a drummer is focused more on timing than anything else, and regardless of how it is presented the latency affects the timing equally. Please note that this explanation is offered without any evidence and is simply a theory.

The keyboardists prefer more latency in the IEM mechanism than the Wedge monitor mechanism.

The keyboardists seem to prefer similar amounts of latency and judge thresholds of quality equally in both mechanisms. The fact that more latency is required in IEM than Wedge to obtain the same threshold implies that similar to the drummers, the keyboardists seem to have a certain amount of latency that affects threshold change regardless of configuration change. This is consistent in the data as well as the low variance of keyboard data. In fact, the amount of difference shows that there is additional latency added to the IEM mechanism to equal the propagation latency in free air for the wedge mechanism. It can be concluded that the keyboard is one instrument where there is a true latency threshold and the change of neither environment nor criteria threshold affects the keyboardist's perception of latency. In addition, the relatively low general sensitivity for the keyboard supports this conclusion.

Vocalists have a high general sensitivity and are thus more likely to change their perception of latency based on variable change.

This is shown in Fig. 39 and implies that the thresholds described in Section 4.3 are not precise in a wide variety of situations.

Drummers and keyboardists have a low general sensitivity and thus are likely to perceive latency very similarly in a wide variety of situations.

This is shown in Fig. 39 and implies that the thresholds described in Section 4.3 are precise and valid in a variety of situations.

6. CONCLUSION

6.1. Summary

This experiment has successfully identified trends in the effects of latency on live sound monitoring. Whereas several trends can only be suggested, a few trends can be concluded from this study. In most cases, more subjects would be needed to get more concrete conclusions. This study introduces many possibilities of future tests to verify and expand on these results.

6.2. Conclusions of the Study

The following are the proven conclusions of the study:

- Sensitivity to latency is more strongly dependent on instrument rather than the individual subject.
- The differences in latency perception from instrument to instrument prevent the ability to define absolute thresholds for the quality of live sound monitoring given a specific amount of latency.
- The Good rating for both mechanisms has a higher threshold than the Fair rating for both mechanisms with the exception of the vocalists.
- If you can reduce the latency in a system to earn a rating of Good in the IEM mechanism, you will receive a rating of at least Good for any other combination of variables; with the exception of the keyboardists whose criticality depends on the wedge mechanism rather than the IEM mechanism.
- The keyboardists and drummers have a latency threshold for a given rating, Good or Fair, which is not very dependent on environmental considerations such as monitoring mechanism.
- The degree of confidence in which you can accept and judge the individual instrument confidence curve graphs in Section 4.1 and the comparison of instruments graph in Section 4.3 is dependent on the instrument's general sensitivity rating, found in Section 4.4. The higher the general sensitivity, the more likely the instrumentalists is to change his or her criteria for latency quality threshold; the lower the general sensitivity the more predictable the instrumentalist's criteria is. Thus the results for instruments with low sensitivity are more accurate and precise that those with

high sensitivity. The following is the order general sensitivities and thus degrees of confidence:

- *Keyboard (High Confidence)*
- *Drums (High Confidence)*
- *Electric Guitar (Moderate Confidence)*
- *Electric Bass (Moderate Confidence)*
- *Vocals (Low Confidence)*
- *Saxophone (Not Confident)*

The actual thresholds of latency perception given monitoring mechanism and grading criteria can be viewed in the comparison of instrument results Section 4.3 by viewing the 85% confidence level graphs shown in Fig. 36 and Table 1. Please note that due to a large variance in the saxophone calculations, these thresholds are not necessarily fully accurate.

6.3. Answers to the Study's Questions

What are the differences in latency perception between two different monitoring situations: Wedge Monitors 4-6ft from the ear and In-Ear Monitors (IEM)?

There are numerous differences in the perception of latency between Wedge Monitors and IEM monitors. The most notable difference is the decreased amount of latency necessary to incite the same quality rating in the IEM with the exception of the Drums and Keyboards. Also note that there is an increase in low latency effects, likely due to the monitoring source being inside of the ear.

What are the differences in latency perception among different instruments? Are these differences statistically significant? Which instruments are more sensitive than others?

The latency perception changes substantially from instrument to instrument. In terms of strict threshold cutoff amplitudes, the sensitivity to quality change based on latency magnitude is roughly as follows:

Saxophone
Vocals
Electric Guitar
Electric Bass
Drums
Keyboard

The degree of confidence in this order is surprisingly almost completely reversed.

Is there a difference between solo delayed monitoring and monitoring one's own delayed instrument while playing with a group of non-delayed musicians?

There seems to be a difference, but only at low latency levels. The difference is not great enough to prohibit the results from being combined to have a more accurate view of general results. Ultimately, a test would have to be designed specifically to answer this question.

How much latency can be present in a signal path before a musician will perceive an artifact in the audio signal?

This is highly dependent on instrument type. Please see Fig. 36 for a worst case estimate and follow the lines for the Good Rating. If we ignore the inconsistent saxophone data, latency values greater than 6.5ms for wedges and greater than 1ms for IEM would likely produce slight artifacts for some instruments.

How much latency can be present in a signal path before a musician will perceive an actual delay in the signal?

This is highly dependent on instrument type. Please see Fig. 36 for a worst case estimate and follow the lines for the Fair Rating. If we ignore the inconsistent saxophone data, latency values greater than 16ms for wedges and greater than 6.5ms for IEM would likely produce some audible delay for some instruments.

6.4. Future Research

This experiment offers many possibilities for further clarifications. Different experimental paradigms are needed to answer questions related to very small latency amounts such as the following:

- Discrimination between an analog and a digital signal path
- Experimentation with regards to the preference of latency in a signal path that is equal to the natural latency in air for the typical instrument's direct sound.
- Differences in latency perception with a non-delayed metronome or musician

In addition, this experiment would benefit from having more subjects. Note that additional experiments may not hold constant variables in the same manner as this experiment allowing different noise factors to be added to the results. In this case the results of the new experiment may not be directly and conclusively comparable to this study's results. Although with careful design, the results could augment the experiment presented in this paper. Lastly, since the most significant variable in latency perception was determined to be the instrument, there would be a large utility in performing the same experimental procedure with additional instruments such as strings or brass to compare the thresholds of more instruments. With more instrument thresholds, a better idea of the lower and upper bounds of latency perception when

considering a random instrument or random population could be suggested.

7. REFERENCES

[1] ITU-R RECOMMENDATION BS.1534-1: *Method for the Subjective Assessment of Intermediate Quality of Coding Systems.*

[2] H. Wallach: "The Precedence Effect in Sound Localization," *Journal of the Audio Engineering Society, Vol. 21 No. 10*, pp. 817-826, Dec 1973.

[3] N. Macmillan and C. Creelman: *Detection Theory: A user's guide*, 2nd edition, Mahwah, NJ: Lawrence Erlbaum and Associates, 2005.

[4] ITU-R RECOMMENDATION BS.1116-1: *Methods for the Subjective Assessment of Small Impairment in Audio Systems Including Multichannel Sound System.*

[5] C. Chafe: "Network Time Delay and Ensemble Accuracy: Effects of Latency, Asymmetry" *Proc. of the AES 117th Conf.*, 2004.