YAMAHA

WHITE PAPER

VOICE –THE MEAT IN VIDEO

Audio quality and user experience generally define the success of a video solution.

By Philip Edholm, President & Principal, PKE Consulting

Much has been written about the advent of the video age. In fact, the use of video conferencing as a business tool is exploding. While much time is spent on the quality of the video: (frames per second, lost packets, pixilation, video codecs, etc.,) there has been less time spent on the value of the audio experience within video conferencing or other business video uses. The focus of this white paper is to understand the value of audio within the video experience and why audio quality and user experience generally defines the success of a video solution.

VIDEO BUSINESS USE CASES

People have been communicating for all of history, from mail and the telegraph to the phone, email and text. Video has long been seen as a watershed transformation in communications. However, video is actually an enhancement to communications. The capability to visualize over a communications or collaboration event has three potential values.

VISUAL FEEDBACK

The most common use of video conferencing is to enable visual feedback between the participants. In a normal business conversation, this visual feedback enables the participants to judge the non-verbal cues that are a critical part of human speech. For example, a nod after a statement means the other party either agrees or at least understands. A participant leaning back and crossing their arms indicates a detachment from the conversation. In customer interactions, seeing responses in a sales situation can enhance the value, and seeing a customer care agent engenders empathy. In all of these cases, video is enabling a set of visual cues that moves the interaction closer to the face-to-face experience.

Group Attention and Interaction

Another common use and value of video is to assure attention and interaction in group settings. While it is easy to “hide” on an audio conference, being on video changes the demeanor and the attention of the participants. If it is easy to multi-task, the participants will naturally reduce attention. When the participants can be viewed on video, there is a natural impetus to be engaged and pay attention.

Data Capture and Transmission

A third use of video, not really about human interaction per se, is to use the video channel to send information. For example, a video stream from a customer to their insurance company can show the damage from an accident, or a video stream from the collision repair shop to the insurance company allows the discussion of the repair cost.

VIDEO VERSUS AUDIO VALUE

In all three of the above examples, the video is part of the media of the overall interactive event. Using currently popular terminology, the audio is the content/information, the video is the context. In the first example, the visual feedback is actually only a confirmation of the audio content, or a context of interpretation. Similarly, in the attention example, the real content of the meeting is in what is being said, while the attention feedback only improves the meeting. In the last example, the video stream is information, but it is actually contextual to the larger discussion about what is causing or driving the need to send the video.

The relationship between video and audio in a video conference is actually very easy to understand if we turn either media flow off. If I turn off the video in a video conference and leave the audio conference in place, the participants can continue relatively unimpeded. They will lose the context of visual feedback or attention, but the core of the interaction event can continue in the audio channel. Conversely, if you stop the audio streaming in a video conference the reaction is dramatically different. Immediately there is much hand waving, followed by texting and finally by a separate voice call. The point is that video does not enable us to communicate content and information in the same way we do through our spoken interactions.

In 1971, Albert Mehrabian published Silent Messages[1], an analysis of how people communicate, both verbal and non-verbal. The research concluded that prospects of a sales process based most of their assessment of credibility on factors other than the words the salesperson spoke – the prospects studied assigned 55 percent of their weight to the speaker's body language and another 38 percent to the tone and music of their voice. They assigned only 7 percent of their credibility assessment 7% to the salesperson's actual words. While this analysis is 38% in a very pejorative area (sales), the underlying truths are critical. While more than 50 percent of the value of a selling 55% type of interaction is visual, nearly 50 percent is also audio (what is said and the intonation of the speaking). While this analysis probably puts more weight on trust factors like body language, the important

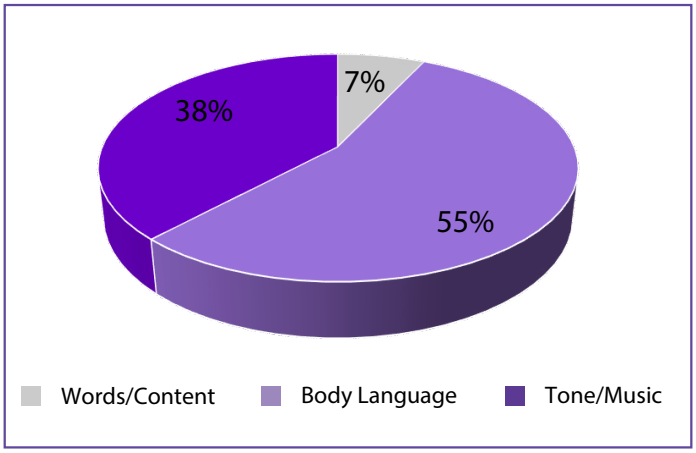


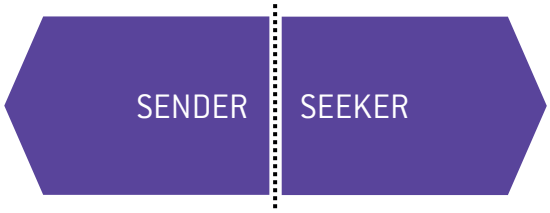
Figure 1.

point is that the audible quality value of a conversation is at least as valuable, if not more valuable than what is said. For us to fully interact, we need to be able to not only see body language (the video component), but also hear the inflections and emphasis that are as important as the words themselves. To do this requires a quality audio experience.

Research has shown that voice is the mechanism humans use to communicate information for task-based activities. In 1974, as part of a study to see how humans would best communicate with a machine, Robert Ochsman and Alfonso Chapanis conducted an experiment to explore which communications mechanisms would best enable a sender of information to communicate with a seeker who needed the information to complete a task[2]. The experiment placed the sender and seeker in separate rooms and the communication interface was changed over many repeated experiment attempts. First they used a teletype (it was 1974 after all), then a handwriting machine, then both, then voice, then voice and others, and finally they opened a curtain on a window between the rooms to simulate video. The results of the experiment should come as no surprise. Adding voice dropped the time to completion by about 70 percent on average, but the fact that "video" did not change the timing indicated a core human characteristic: humans solve problems and complete tasks by working together and TALKING. Even the old adage, "A picture is worth a thousand words," shows how speech is critical. A picture of four people, dressed in business casual attire on a beach with Diamond Head in the background says much. They are probably business people, going out on the beach in Hawaii. But that is all the picture says. Who is the third person from the right? Why are they there?

Visual video information is actually the icing on the audio cake that is the foundation of any video conferencing event. While video adds context and feedback and empathy, it can only contribute to the success of the event in an additive way. For

> *"Audio must be capable of providing the content and information so the video context can have value."*



- Typing
- Handwriting (electromechanical)
- Handwriting and Typing
- Handwriting and Image
- Voice
- Voice and Typing
- Voice and Handwriting
- Video and Voice
- Communications (FtF)

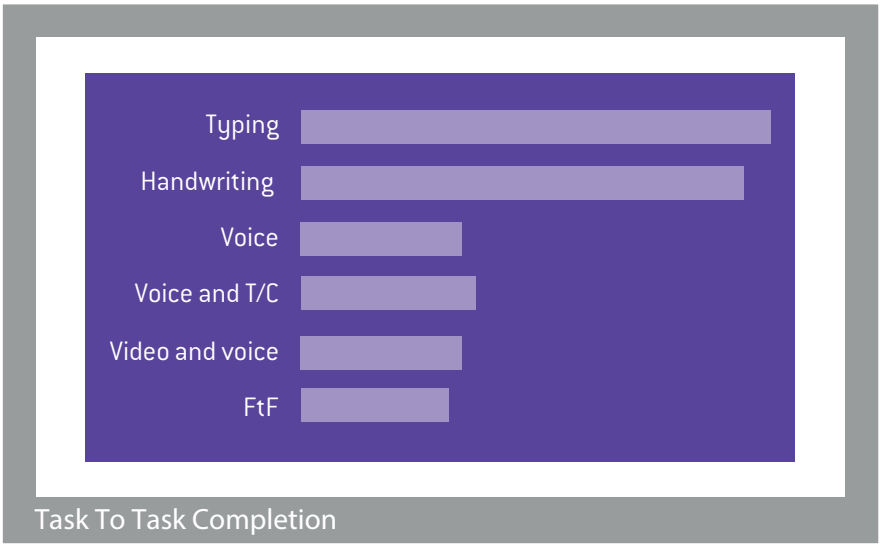


Figure 2.

the event to succeed, the underlying voice and audio must be capable of providing the content and information so the video context can have value.

TASK VERSUS SELLING COLLABORATION

There are two types of communications events, selling collaboration and task collaboration. Selling collaboration is defined by one or more of the participants selling a concept, product, idea or other value to the other participants in the event. In selling collaboration, visual feedback is critical. Assuring that the other parties hear the speaker, understand the point, and may or may not agree is critical. If a point is not understood or agreed to, building the next steps in a logical argument will fail as the foundation is not well constructed. Similarly, visual feedback is a great way of gauging agreement or commitment to an idea or activity.

Task collaboration is more focused on the actual information stream rather than the visual feedback. In a task collaboration, both parties have similar motivations to complete the task, rendering the need for visual feedback moot. However, task-based collaboration often requires a relationship between the participants. Shared values and culture make task-based collaboration easier. In modern organizations where the parties are geographically dispersed, part of an ad hoc team or from different cultures, the potential for task-based collaboration to go awry is high. A good example of this was the cockpit environment in Asiana Airlines Flight 214 immediately before the crash in San Francisco in 2013. The conclusion of the NTSB investigation was that the crew mismanaged the landing glide pattern, at least partially because the junior pilots who sensed a problem did not speak up to the actual pilot. This is a scenario similar to that which video visual feedback can solve. If a participant looks uncomfortable with an option it can be seen on video, while it is not noticeable in audio.

For all of these reasons, video conferencing is exploding in use. For example, on Skype, about 40 percent of the calls are video calls. While many of these are personal and familial, it is a clear indicator that video, when available for little or no added cost over voice, combined with an easy-to-use paradigm will be used and adopted. The result is that, in organizations that have enabled video, the adoption and use very rapidly approaches that 40 percent or even higher level. As users get comfortable with video, they use it to accelerate discussions and reduce time to decisions.

> *"Make audio that enhances the video experience, not detracts from it."*

AUDIO IS CRITICAL TO VIDEO CONFERENCING

As video solutions take off, the audio component often gets ignored. While there is an initial infatuation with the video, the lack of quality audio often impacts the conference. This can show up in multiple ways. For example, many smaller rooms use a simple “puck”-type USB microphone/speaker. These devices have relatively narrow microphone range and low audio quality. The result is that often the other participants in the session cannot hear all of the speakers in the room leading to the continual comments to move closer to the mic or to pass the puck. And in the room, the quality of audio varies for the attendees depending on position, distance from the speaker, etc.

In many organizations the result is that the telephone in the room is used for voice/audio as an independent call, separate from but simultaneously with the video system. As most video conferencing systems enable telephone dial-in, the participants mute the in-video audio and use the telephone. While this may result in marginally better audio, it has a significant impact in exposing the latency in video and results in significant lip sync delay between the video and audio.

The challenge is to make audio that enhances the video experience, not detracts from it. To do this, it is essential to take advantage of the wideband codec capabilities that are included in the built-in audio of video conferencing systems. This enables the full range of speech frequencies as well as the dynamic range. See the Revolabs white paper “Audio Quality Matters” for a complete description of modern voice audio capabilities.

This is critical to ascertain the subtleties that Mehrabian noted in Silent Messages. As he indicated, 38 percent of the impact of an interaction is the tone and musical flow of the speaker. As most video conferences involve some form of selling, the ability to carry this information in the voice/audio path is critical. One of the reasons many users go back to just a telephone call is the lack of good audio/voice quality in their video experience.

THE CHALLENGES WITH CONFERENCE ROOMS

In traditional large video conferencing rooms, the audio has been delivered by complex systems with multiple microphones (either desk-mounted or hanging from the ceiling) combined with sophisticated control systems including selective muting and other features. While these systems, that often cost thousands of dollars for the audio system alone, are effective, they do not translate well to the smaller room. In the smaller room a simpler and lower cost option is required.

The audio issues we experience in rooms are a function of room shape, room acoustics, distance between the people speaking and the equipment, and other factors. To deliver quality audio in tandem with video in a room requires a microphone and speaker system that can both capture the audio and play it back at volume levels and clarity that can be understood. A good example is the microphone. In a typical speakerphone there is only one microphone, as it is assumed that the person speaking is in front of the phone. However, in a conference room, there will be multiple participants, typically sitting around a table. A single microphone is not a reasonable solution. If it is placed on the device so it is omnidirectional (i.e. facing up), the off-axis response will create variance in the relative volume and tonality of the talkers. For this reason, most conference room systems have multiple microphones in an array that pick up audio from all around the device. However, this introduces the immediate challenge of managing the audio between the microphones for clarity. For example, if there are four microphones on the four corners of the device, they must be engineered so one voice is not picked up by multiple microphones at different delay times. If not built properly, the audio quality of multiple microphones can actually be worse than a single microphone.

Another critical value is the ability to suppress echoes. In communications systems, echo introduced in an endpoint generally does not impact the people at that endpoint, but the other participants. In other words, if I am talking and your device is coupling some of my voice sound back into your microphone and sending it back to me, I hear my own voice delayed by the latency of the video system, often over 200 msecs. If this echo is at sufficient volume it can make the audio channel unusable. However, even at a low volume it can be disconcerting and make communications difficult. In a conference room environment, echo becomes quite complex. While the primary echo of the speaker coupled into the microphone is relatively easy to manage, when there are multiple microphones to manage, combined with various reflective surfaces, the actual echo pattern can be much more complex.

Through all of this the capture of the voice/audio in the room is critical. Choosing a video room solution that does not adequately accommodate the audio part of the video conference will dramatically impact usability as well as adoption. The goal of the conference room audio solution is to enhance the video experience, not detract from it.

Simple Video – Advanced Audio

As the balance of investment between video and audio is considered, it must be done with due consideration of the actual usage of the room. For example, in a smaller room such as a huddle room, the need for advanced camera features like pan/tilt and zoom are not necessary. In this case, a simple camera will suffice for the video meetings. By optimizing the spend on the camera, an audio device can be acquired that delivers quality audio to enhance the meetings that will be held. Rather than buying an expensive 4K monitor that has a quality level that most video conferencing systems will not approach, settle for 2K video technology and invest in the audio components.

The outcome of focusing the investments will be better meetings. This is especially true as video conferencing moves out of the boardroom and into the myriad of smaller spaces. Estimates by IDC and Gartner put the percentage of conference rooms that have video installed at 4-7 percent. For the other 94 percent of conference rooms that do not have video, the inclusion of video is for a full range of uses, including selling activities, but also coordination, attention and commitment. As the systems in conference rooms will be used for many task-based activities where the video is secondary to the audio or not even used at all, the audio/voice part of the room is the critical investment.

USB Camera – Better Audio

One of the greatest bargains in peripherals today is the low cost USB camera. These cameras, employing sophisticated sensors and on-board codecs, can provide very high quality video at relatively low cost. A number of emerging video endpoints are moving to using these cameras for smaller rooms. If the pan/tilt or zoom features of more expensive cameras are not required, a simple USB camera can suffice.

A good example of this is Microsoft Project Rigel. Rigel is a program in Microsoft to enable a video room endpoint using a simple processor box with Windows 10 and a Skype for Business room experience application. The video is provided by a USB camera. This can be a simple camera or one with more complexity, but the assumption is that most rooms will have a simple camera. As in the desktop configuration, most cameras also have a microphone, but, as discussed above, these single microphones have significant issues and the audio is generally driven through the speakers in the flat panel display, most of which are no longer designed for quality audio as most home users augment their television sound with a dedicated home theater system.

For most smaller rooms, the ideal solution is a basic video system with a high quality low cost USB camera combined with an audio device that assures that the conversations will be optimized, both for the in-room participants as well as those remote. The key advantage to this strategy is that focusing on good audio quality typically means the audio device can be used through migrations. While video continues to improve with new technologies like 4K, advanced codecs like h.265 or VP9, a quality audio investment can continue to deliver value. Whether the video system remains with the initial vendor, or there are changes as cameras and codecs are replaced or upgraded, a quality audio investment can be used throughout the changes and video migrations with any system.

The Multi-Purpose Room

Another significant consideration is whether the conference room audio device is used just for video conferencing or for other events like audio conferences and web collaboration. Choosing a simple USB-connected puck will limit that device to video. Advanced conference room audio devices for video often include or have optional configurations that enable them to be VoIP phones or provide audio for in-room PCs. This makes the device multi-functional, minimizing in-room clutter and assuring high audio quality for all events.

OPTIMIZING YOUR VIDEO CONFERENCE ROOM INVESTMENTS

While video conferencing is often the driver for upgrading conference rooms and fitting them with new equipment, the audio strategy should be front and center. In many situations the audio considerations may actually outweigh the video. This is clear if you go into virtually any boardroom that includes a video conferencing system. The video system is generally relatively simple, a camera designed to capture the room (perhaps two with some of the newer voice tracking systems), combined with a video display, either a flat panel or using the projector. In many ways, this is virtually identical to the video components in any other room. However, a look at the audio will show dramatically different investment. There will typically be a microphone for every two or three seats around the board table, either table-mounted or ceiling. The system will always include muting capabilities.

However, the microphones are the tip of the iceberg. In addition to the array of microphones and multiple speakers, there will be a very sophisticated processing system that is optimizing the environment to assure that conversations can be heard and understood. In fact, the audio system in a board room often exceeds the cost of all of the other components combined. As organizations move to include video in the full range of meetings and conference spaces, the same investment strategy should be applied. In any conference space, investing to assure that the audio and voice quality is absolutely best in class is critical. In fact, while video systems can be upgraded with new displays, choosing the right audio component from day one is the most critical element of the overall solution.

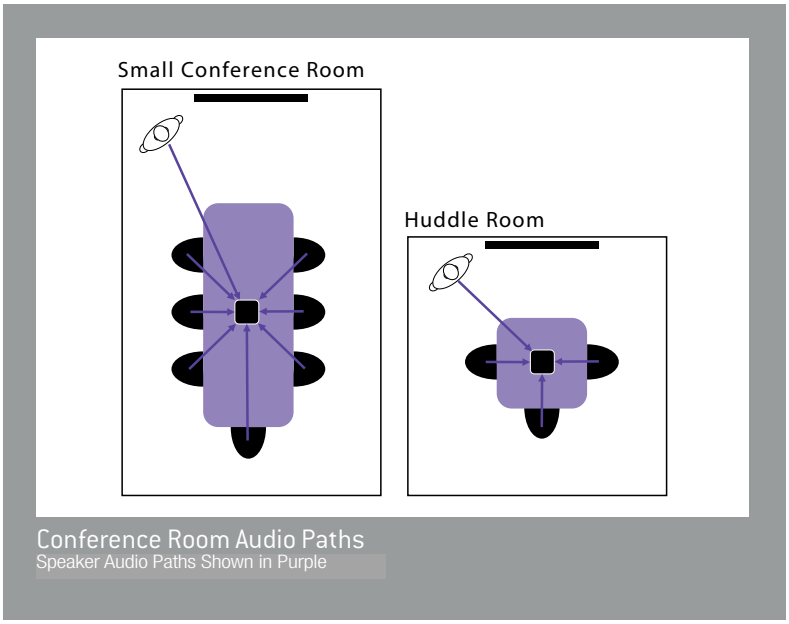


Figure 3.

CHOOSING THE RIGHT AUDIO DEVICE/SYSTEM

If you have read this far, you are probably convinced that audio is a, if not the, critical component of a quality video experience. As you evaluate options for audio devices for your conference rooms, there are several critical capabilities and features to include.

Microphones

The device must have the capability to receive audio from 360 degrees in the room. Figure 3 shows a typical small conference room and a huddle room. The dark blue lines show the audio paths from each talker to the audio device on the conference table. There are two critical points that must be noted. First, in virtually all rooms, sound will arrive from the 360-degree arc. Second, the distance from the audio device can vary greatly. This is especially true when there is a front-of-room presenter. As can be seen in the figure, a single microphone device will have a hard time managing the multiple talkers.

The key to analyzing the performance of an audio device is to use it in a conference and walk to the edges of the space during the call and have the person on the other end indicate if they are hearing a change in volume or clarity. The goal is

to have the audio that is received by the remote user not change. A second test is to measure if the system is managing microphone sensitivity to the speakers. To test this, have someone speak and then have a lower level noise on the other side of the room. Sophisticated systems will block out the lower level sounds, enabling remote users to focus on the actual speaker.

A final microphone capability is to determine whether the system will allow the connecting of a remote microphone for larger environments. This is advantageous when the table exceeds 6-8 feet; the talker's distance does not become an issue for quiet talkers who are located away from the primary unit and microphones.

Speakers

With modern audio for video conferencing using wideband codecs, the sound quality often approaches or exceeds the quality of music CDs. Just as in music and singing, the quality of the speakers now has a major impact on the overall sound. With wideband audio codecs, the dynamic range of the speaker is now captured. As this is a critical part of the "Silent Message," being able to convey those subtle variations in tone and volume are critical. This requires a speaker that has both a wide frequency range, as well as the capability to deal with the wider dynamic range enabled.

There are two ways to test an audio device. The first is to do actual conferences with the remote participant on a high quality microphone such as those in high end headsets. Have the remote speaker talk though a range of emotions and volumes and see if you can sense the tone and musicality. A second test is to connect the system to a PC (typically these units have a USB connection) and play a high quality musical selection. In many ways, the ability of a device to do a reasonable job of music reproduction is a great indicator of the ability to convey the subtlety in human communications.

Echo and other Audio Processing

To test the echo and other processing, the best way is to initiate a conference session to the device with a system that is assured to be using a high quality wide band audio codec and operation in full duplex mode. In full duplex mode you can be talking and hear someone from the other end talking at the same time. A good example of this is a system using the WebRTC Opus codec, while many video systems do this as well. Connect to the room systems from a remote

location using a high quality headset. For the room environment, choose a room that has hard walls (glass, white boards) that increase the in-room reflections and will increase the challenges for the echo processing. While talking on your headset, listen carefully to hear if any remnants of your speaking can be heard in the headset. If you hear yourself, no matter how faintly, you are hearing echo. One way to test for echo is to talk in short bursts of about 100 msecs so you can clearly hear the echo if it comes.

VoIP Phone and PC Audio Connections

The last factor to consider is whether the audio device can also function as an IP phone to your VoIP PBX or Unified Communications system, as well as the ability to use the audio device for simple in-room audio from a PC, Mac, smarphone or tablet. Some vendors like Revolabs offer different versions of their products that include the VoIP phone capabilities as an option.

Similarly, the device should enable users to connect their own devices, either through USB, Wi-Fi, or Bluetooth to the device to play audio for an in-room experience or as part of sharing with the video system. This allows the conference room to have the full advantage of enhanced audio quality for all events.

CONCLUSION

The need for quality audio in all types of conferences and meetings is clear. In video conferencing, quality audio is not only desirable, but is, in fact, the primary limiting factor in the success of most events. Having a clear audio strategy is critical for both successful use and adoption of video conferencing systems.

Investing in quality audio for video conferencing rooms should not be an afterthought, but should be a core part of the overall decision process. Choosing quality audio first is the best way to assure the overall acceptance and use of your new video solution as audio and audio quality represents at least 50 percent of the value in every video conference. Don't let it become the limiting factor in your investment.

References:

1 Albert Mehrabian (1971), Silent Messages: Implicit Communication of Emotions and Attitudes", Wadsworth Publishing Company

2 Ochsman, Robert B. and Chapanis, A. (1974). "The effects of 10 communication modes on the behavior of teams during co-operative problem-solving." International Journal of Man-Machine Studies 6: pgs. 579-619.

SHARING PASSION & PERFORMANCE

Yamaha Unified Communications, Inc.
144 North Road #3250
Sudbury, MA 01776 USA
+1 800-326-1088
uc-sales@music.yamaha.com

Yamaha UC EMEA
190 High Street Tonbridge, Kent
TN9 1BE, UK
+44 (0)1732 366535
uc-salesemea@music.yamaha.com

Yamaha UC APAC
+852.8108.8820
uc-salesapac@music.yamaha.com