Yamaha's High-quality Sound Technologies Which Further Accelerate Remote Communications

Issued in March, 2015

Foreword

Recent years have ushered in a genuine unified communications era. Advances in not only hardware based teleconference systems, but in PC software based web conference systems as well have elevated their performance, and along with the HD-like high definition images, an even higher quality of sound came to be sought. Yamaha had been looking ahead to elevating the sound quality of those web conferences, and has developed a variety of sound signal processing technologies such as an adaptive echo canceller which is effective up to 20 kHz. Along with the excellent operability which can be simply used by anyone, Yamaha is going to provide high-performance devices which are sufficiently compatible for the next generation of high-quality sound web conferences.

Table of Contents

Purpose of This White Paper

Yamaha has developed the YVC (Yamaha Voice Communication) series to update our company's heretofore concepts and technologies for the unified communications era. In this white paper, in conjunction with introducing the core technologies of the very first model the YVC-1000 and providing commentary on it, we have performed comparison tests that covered a specific sampling of systems (subsequently called, "commercial conventional systems") which had been selected from conventional popular type models, and the findings verify that the product, the YVC-1000 facilitates much more pleasant and smoother teleconferencing via the more advanced technologies.

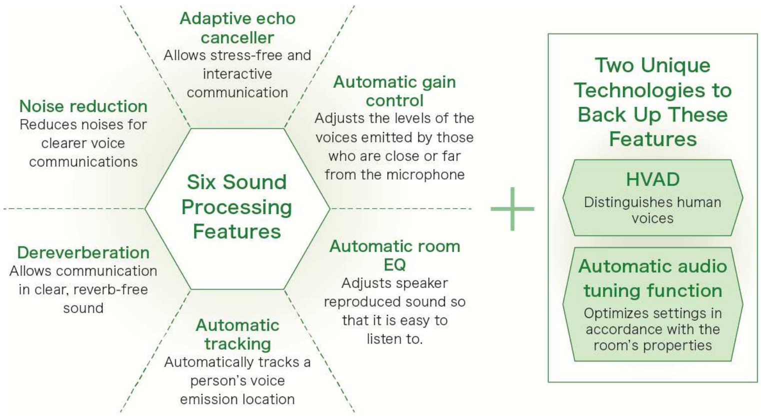
Term and Definition

Term	Definition
DUT: device under test	Microphone & speaker system or microphone under test.
RT: reverberation time	A period of time until reverberation is reduced by 60 dB.
FE: far end	Intended party location in remote communication.
NE: near end	One's own location in remote communication.
Normal microphone	Microphone without signal processing. In this white paper in particular, unidirectional boundary microphones are used.

Chapter 1 Overview of YVC-1000 Sound Processing

The YVC-1000's sound signal processing technologies actualize high-quality sound and pleasant communications via the six basic signal processing technologies in the diagram below, which are configured by two of Yamaha's unique technologies that support them.

(Figure 1-1)



The details of signal processing features are described in chapter 2. This chapter describes the overview and configuration of these features and the outlined image of their correlations.

1. Signal processing technology to improve microphone sound-pickup quality

The Adaptive echo canceller, Automatic tracking, Noise reduction, Dereverberation and Automatic gain control features correspond to this technology. The high-accuracy human voice distinguishing technology of "HVAD - Human Voice Activity Detection (See Chapter 3)" that Yamaha has developed on its own for the YVC series is installed as a function to enhance some of these features.

2. Signal processing technology to improve speaker sound reproducing quality

Automatic room EQ is a signal processing technology used for acoustic signal reproduction. It automatically controls the built-in speaker equalizer depending on the environment currently used to improve the listenability of reproduced sounds.

3. Signal processing technology to automatically optimize acoustic settings

YVC-1000 is equipped with the automatic audio tuning function (See Chapter 3) as a Yamaha's unique technology. YVC-1000 automatically optimizes its acoustic settings by learning about the acoustic environment of a room where it operates. Prior to its use, simply press the tuning fork button to activate the automatic audio tuning function, which immediately allows the acoustic settings to be optimized in advance. Implementing this function optimizes the parameter settings of the signal processing technologies to make full use of the capabilities of the individual technologies. Also, this function automatically compensates the differentials in delay times or frequency responses between the built-in and external speakers, if any, to improve the listenability of reproduced sounds.

Chapter 2 Roles and Capabilities of the Six Sound Processing Technologies

1. Adaptive Echo Canceller

1.1 What is the adaptive echo canceller?

The adaptive echo canceller is a function which eliminates the sounds on the microphone side that are generated from the speakers, and the echoes that the microphone picks up.

1.2 Superiority of YVC-1000

Although the fundamental performance of echo cancellers has been generally established, YVC-1000 is superior to commercial conventional systems in sound quality in rooms with long reverberation times and during interactive simultaneous communication within the HD band (100 Hz - 20 kHz).

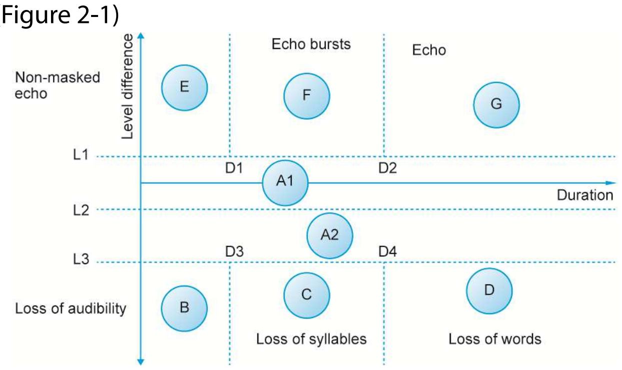
Evaluation method:

As a method to objectively evaluate the performance of echo cancellers, Yamaha adopted the "Tdos S4-AHQ052" method (Figure 2-1 and Table 2-1) proposed by ETSI (the European Telecommunications Standards Institute) and thereby evaluated the echoes and the degree of losses in inserted sounds.

This method evaluates the echo canceller's performance by comparing each time that the sound levels correspond when the same near-end sound is picked up via the microphone's sound pickup (the first party) without speaker reproduction, and the microphone's sound pickup (the second party) with speaker reproduction. If the sound level of the second party is larger than that of the first party, it indicates that an echo is generated. If the sound level of the second party is less than that of the first party, it indicates that an insertion loss (loss of sound volume) is generated. An echo canceller is more excellent in performance as the absolute value of echoes and insertion losses is small and their duration time is short.

ETSI evaluation model

(Figure 2-1)



(Table 2-1)

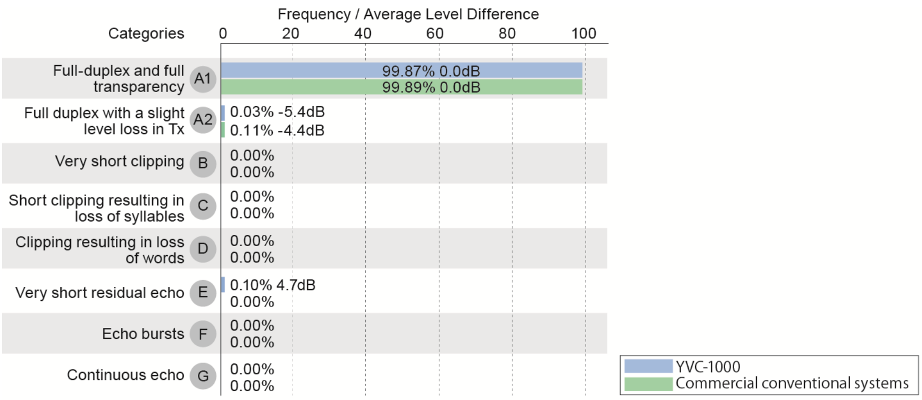
A1	Communicates without loss of information.
A2	Communicates without loss of information, but turns down the volume.
B	Communicates information, but may omit information for a short time.
C	Communicates information, but may omit any syllables.
D	May omit any words.
E	Communicates information, but generates echoes for a short time.
F	Communicates information, but generates echoes.
G	Connects while generating echoes.

L1	4 dB
L2	-4 dB
L3	-15dB
D1	25 ms
D2	150 ms
D3	25 ms
D4	150 ms

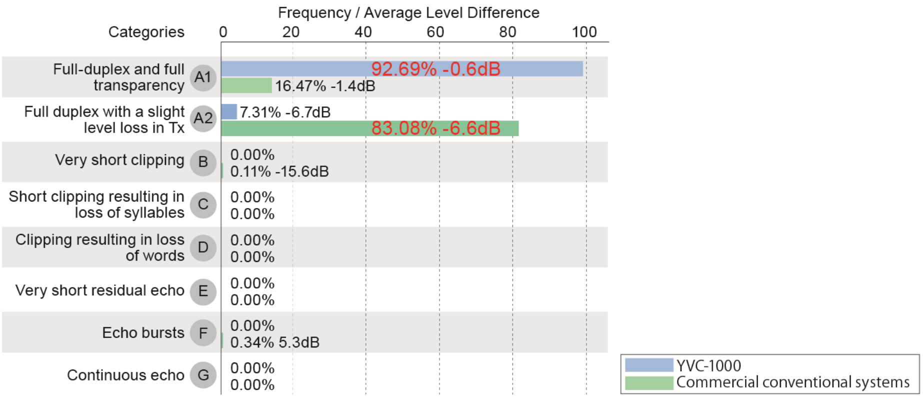
Applying measured data results to two indexes of L (echo for plus, loss for minus) and D (duration time of echo or loss) in the right table of Figure 2-1 derives the appropriate evaluation sections (A1 - G) in the left table.
Finally A1 indicates the highest performance of the echo canceller. In view of the insertion loss, its performance is degraded in the sequence of A2, B, C and D. In view of the echo, its performance is degraded in the sequence of E, F and G.
Moreover, the performance of an echo canceller is evaluated by measuring the frequency of occurrence of A1 to G and the mean value of level differences at the times of their occurrences.

< Comparisons with commercial conventional systems and validations >
Validation results:
The following Graphs 2-1 and 2-2 show the measurement results that are applied to the ETSI evaluation model.

(Graph 2-1)
Under the condition of one-way communication from the far end, both models were classified and evaluated into the A1 category, thus sufficiently cancelling echoes.



(Graph 2-2)
When both parties are simultaneously conversing, the commercial conventional systems their A2 frequency increases, and the sound levels at that time drop on average -6.6 dB. In short, the volume is halved in 83% of a portion of the conversation which is similar to the sound degradation that is occurring. On the other hand, with the YVC-1000 the A1 frequency is overwhelmingly higher than that of commercial conventional systems, and the highest in conversation quality which is free of echoes and insertion losses is confirmed.

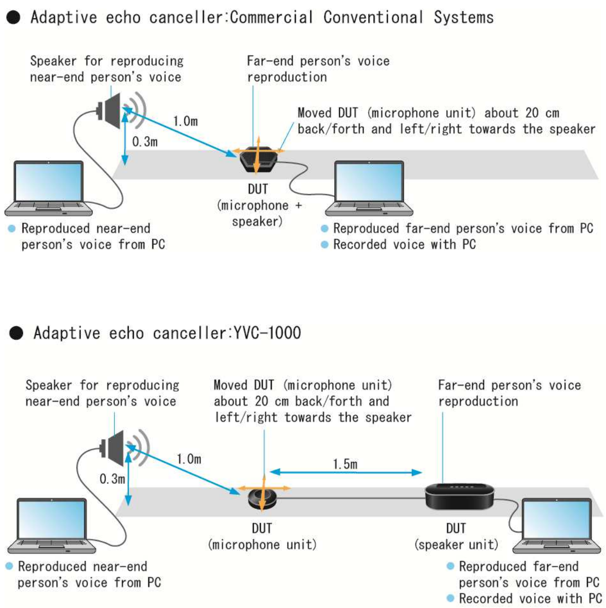


Data measuring method:

In order to simulate the above communication image, the measuring was performed with the system configuration in Figure 2-2.

* The above validation results were obtained using YVC-1000 firmware Ver.1.05.

(Figure 2-2)



2. Noise Reduction

2.1 What is noise reduction?

Noise reduction is a function which automatically detects steady noises such as those from air-conditioners and PC fans that are picked up by the microphones and eliminates just those elements from the sound pickup signals. Herewith, even if noise sources are present, they can be turned into clear sound pickups.

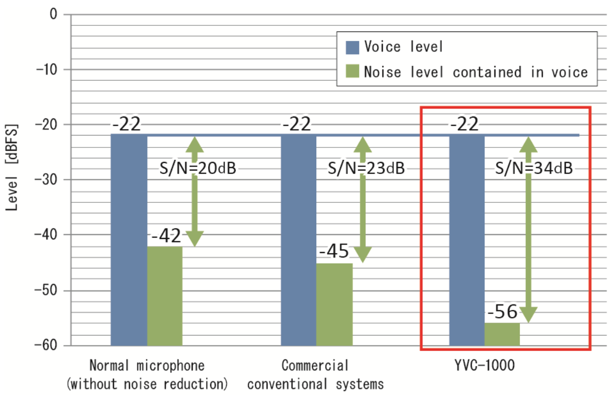
2.2 Superiority of YVC-1000

The commercial conventional systems eliminate noise only when there is no voice. However, they are unable to eliminate noises which coexist with voices. If this is the case, the predecessor generally eliminates steady human voices and music sounds by regarding them as noises. In contrast, YVC-1000 can eliminate noises even if they coexist with voices. In other words, with YVC-1000, HVAD is combined into the evaluating of steady sounds, thus misidentified eliminating such as that by commercial conventional systems is averted, so that just noises can be eliminated.

< Comparisons with commercial conventional systems and validations >

Validation results:

(Graph 2-3)



Graph 2-3 indicates the measuring of noise reduction capabilities in environments with background noise levels of 48 dB (air conditioners set to high).

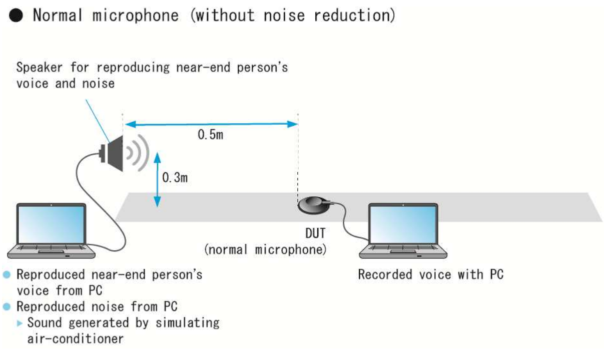
With commercial conventional systems it is known that background noise levels hardly drop in the voice zones. Because of that, the real S/N ratio is 23 dB, and for conventional microphones without noise reduction, the result is that there is not a large difference. On the other hand, since the YVC-1000 significantly improves the S/N ratio, the sound quality becomes equivalent to that of a quiet room.

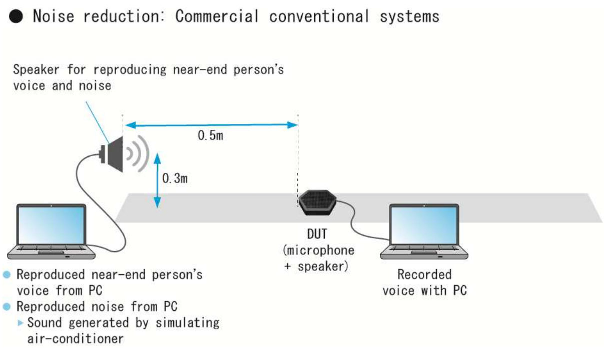
Data measuring method:

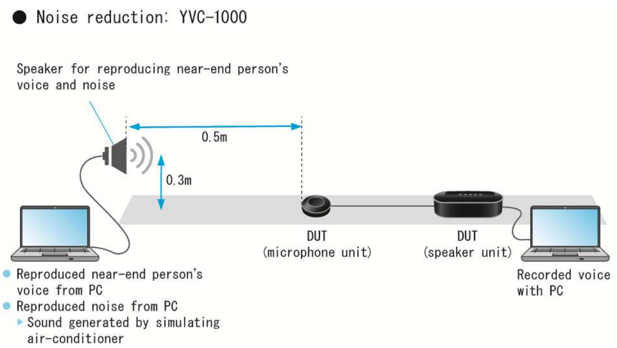
In the configurations as in Figure 2-3, an S/N ratio (signal-to-noise ratio) was measured for comparison with a normal microphone (without noise reduction), YVC-1000 and a commercial conventional system.

* The above validation results were obtained using YVC-1000 firmware Ver.1.05.

(Figure 2-3)







3. Automatic Gain Control

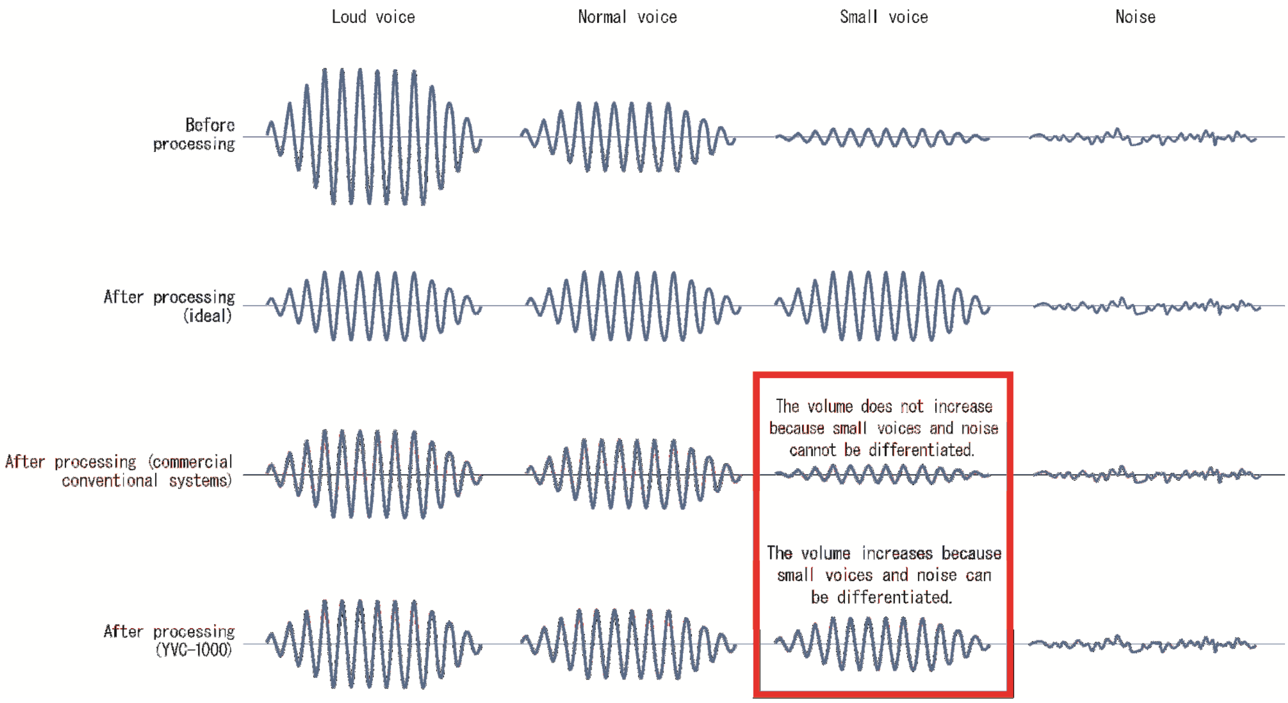
3.1 What is automatic gain control?

Automatic gain control is a function for adjusting the voices to be sent from the communicating parties to a constant level by automatically controlling the gain depending on the level of individual picked-up voices. Additionally, this function is required to suppress unnecessary sounds such as noises and unusual sounds. This allows the communicating parties to hear stress-free sounds.

3.2 Superiority of YVC-1000

With commercial conventional systems since it is difficult to distinguish noises and low voices, it is hard to raise the volume of low voices. In regard to that, since YVC-1000 has raised the accuracy of voice determinations via HVAD, sounds can be firmly raised without low voices and noises being determined.

(Figure 2-4)



< Comparisons with commercial conventional systems and validations >

Validation results:

(Table 2-2)

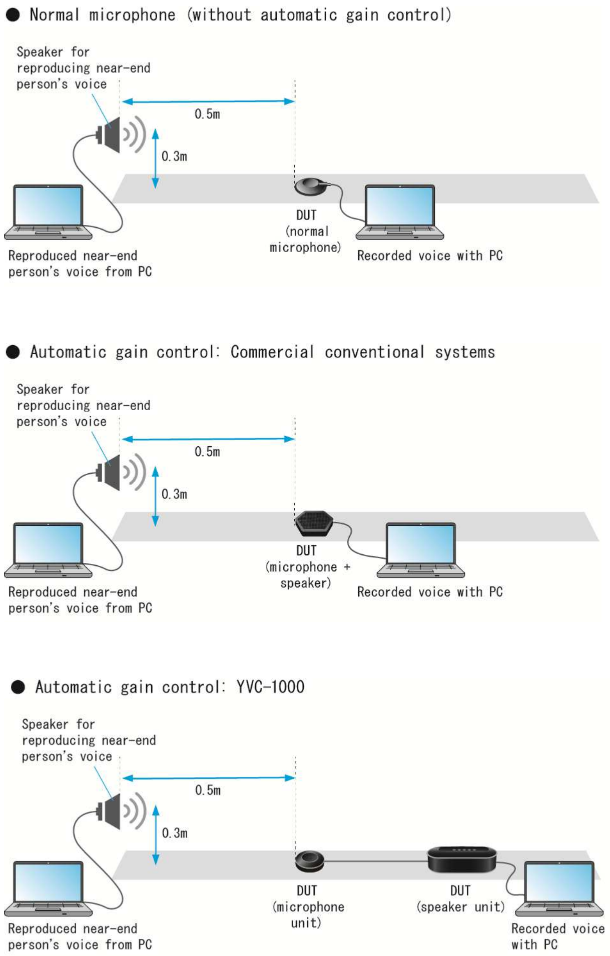
	Normal microphone (without automatic gain control)	Commercial conventional system	YVC-1000
Zone 1 (normal voice)	0 dB	0 dB	0 dB
Zone 2 (loud voice)	+8 dB	+4 dB	+4 dB
Zone 3 (keyboard typing)	−21 dB	−24 dB	−23 dB
Zone 4 (small voice)	−9 dB	−8 dB	−4 dB

Table 2-2 lists the results of the comparative measuring of pickup sound levels in individual conditions (sections). In this table the sound pickup levels in the normal volume zone are designated as the reference (0 dB) and individual pickup sound levels in other sections are expressed relatively. As indicated in the table, what is especially being emphasized in actual conferences, is in the performance whereas the low voices are drawn out, which confirms that there is superiority over the +4 dB of commercial conventional systems.

Data measuring method:

In the configurations in Figure 2-5, the sound pickup levels were measured under each of the conditions in Table 2-2 via normal microphones (without automatic gain control), and YVC-1000, via commercial conventional systems.
* The above validation results were obtained using YVC-1000 firmware Ver.1.05.

(Figure 2-5)



4. Automatic Tracking

4.1 What is automatic tracking?

Automatic tracking is a function whereas clear sound pickup is implemented that automatically tracks the speaker within a room, and focuses in on that voice. In addition, since it pick up the sounds of just human voices without responding to noises, it demonstrates its effectiveness when it is utilized in conferences held in noisy rooms with lots of people.

4.2 Superiority of YVC-1000

YVC-1000 picks up voices emitted from a person's location that has been detected using the microphones array control function. The accuracy of detecting a person's voice location is dramatically improved by the HVAD technology. (Refer to 1) of 1.2 in Chapter 3.) Therefore, even if there are the sounds of papers being rifled through or loud sounds on the conference tables, the focus of the microphones will not be switched as long as they are not the voices of humans.

< Comparisons with commercial conventional systems and validations >

Validation results:

(Table 2-3)

	Voice 1	Keyboard noise	Voice 2	Papers being rifled through noise	Voice 3
Commercial conventional system	-27.0 dB	-40.8 dB	-26.2 dB	-30.4 dB	-25.9 dB
YVC-1000	-27.0 dB	-55.3 dB	-26.2 dB	-49.8 dB	-25.5 dB
Difference	0.0 dB	14.5 dB	0.0 dB	19.4 dB	-0.4 dB

As indicated in the table, in comparison with commercial conventional systems the sound pickup level differentials are notable when there is noise insertion. In the keyboard noise zone, for example, YVC-1000 provided a sound pickup level of -55.3 dB with respect to -40.8 dB for the commercial conventional systems, thus it suppresses the noise levels by 14.5 dB more than conventional systems do. In regard to the commercial conventional systems responding and focusing on the noise orientations, with the YVC-1000 noises and voices are distinguished via HVAD, and the performance differentials are that the noise orientations are not focused on.

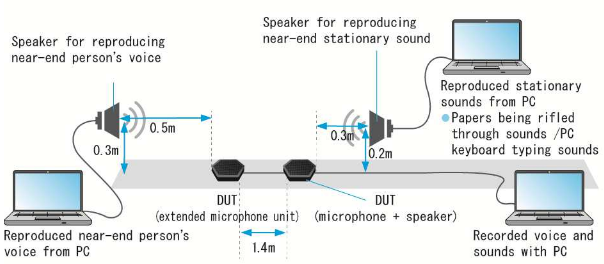
Data measuring method:

In the system configurations in Figure 2-6, the accuracy of detecting a person's voice location was verified. The performance differentials were compared between the YVC-1000 and commercial conventional systems as to whether the voices and noises were alternately reproduced, and whether the noises and voices were misidentified.

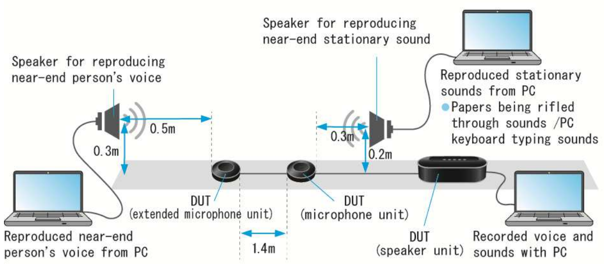
* The above validation results were obtained using YVC-1000 firmware Ver.1.05.

(Figure 2-6)

● Automatic tracking: Commercial conventional system



● Automatic tracking: YVC-1000



5. Dereverberation

5.1 What is dereverberation?

Dereverberation is a function to make human voices clear by eliminating the components that adversely affect vocal clarity (listenability) even if this system is used in a room subject to excessive natural reverb.

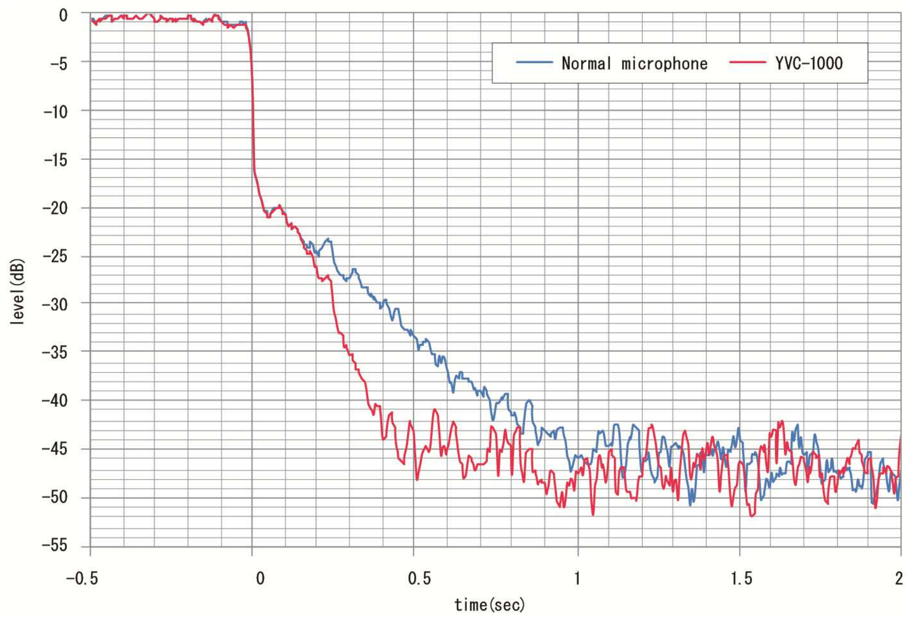
5.2 Superiority of YVC-1000

In the sounds that are picked up by the microphones, the component levels that have determined the reverberation components are dropped. Since the reverberation component frequencies and levels change moment by moment, YVC-1000 responds to those changes and appropriately makes determinations. This kind of function is scarcely equipped in commercial conventional systems.

< Performance validation >

Validation results:

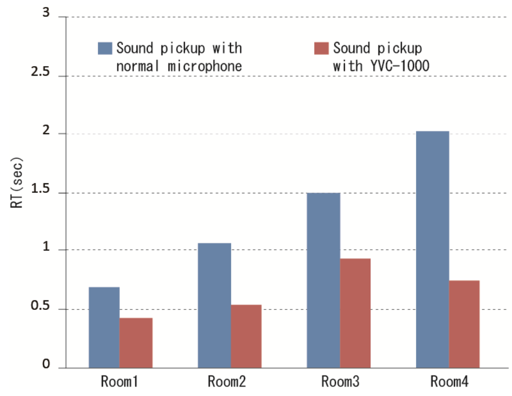
(Graph 2-4)



As shown in Graph 2-4, a measuring sound (white noise) was reproduced from a supposed person's location and then the reverberation convergence time from the reproduction stop time (0 sec in the graph) was compared between a normal microphone (without dereverberation) and YVC-1000's microphone. It is clear from the results that YVC-1000's reverberation converges more quickly than the normal microphone's.

As additionally shown in Graph 2-5, each reverberation time (RT) in a normal microphone and YVC-1000 was measured in four rooms that have different reverberant properties.

(Graph 2-5)



Room 1	RT 0.6 sec (equivalent to a small conference room)
Room 2	RT 1.0 sec (equivalent to a lecture room)
Room 3	RT 1.5 sec (equivalent to a multipurpose hall)
Room 4	RT 2.0 sec (equivalent to a concert hall)

There are some differences depending upon the reverberation properties in the rooms, thus it is known that basically the reverberation times can be shortened by up to half.

(Table 2-4)

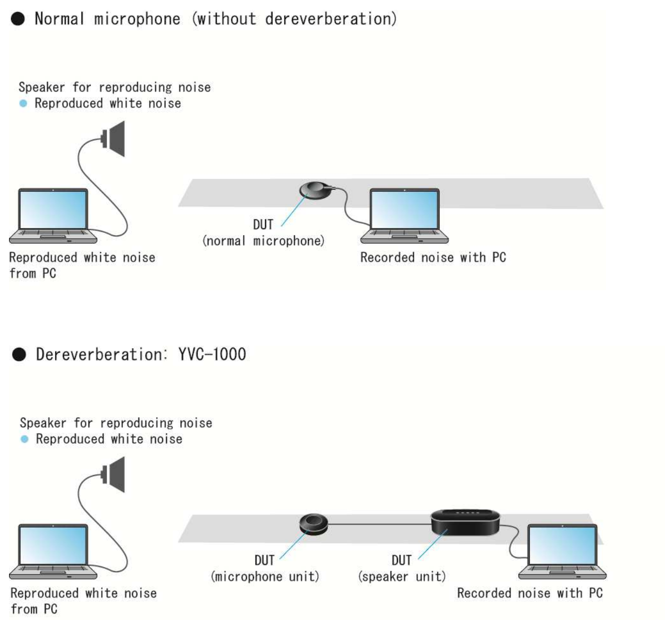
Reverberation time [sec] (without DeReverberation)	2.032 sec
Reverberation time [sec]] (with DeReverberation)	0.752 sec

Data measuring method:

In the configurations in Figure 2-7, the reverberation time calculated from the pickup signals was measured with a normal microphone (without dereverberation) and YVC-1000.

* The above validation results were obtained using YVC-1000 firmware Ver.1.05.

(Figure 2-7)



6. Automatic Room EQ

6.1 What is automatic room EQ?

The automatic room EQ is a function which automatically responds to the acoustic properties of a room and applies equalization to the speakers to facilitate the listening of reproduced sounds. YVC-1000 uses voice signals during conversations, and can even make real-time and automatic adjustments, however the automatic acoustic adjustment functions (refer to 2) of 2.2 in Chapter 3) can also be implemented, so the settings can be optimized in advance prior to the start of conversations.

6.2 Superiority of YVC-1000

The YVC-1000 automatic room EQ has two mechanisms. One is a mechanism that cuts the superfluous upsurge in low sounds, and the other one is a mechanism that cuts the large band frequency of the reverberation. Since the superfluous upsurge in low sounds might impede (masking) the listening of midrange sounds, the listenability of voices is improved by cutting superfluous low sounds. In addition, since even superfluous reverberation components impede (masking) the listening of trailing voices, the listenability of voices is improved by cutting the large band frequency of reverberations. Commercial conventional systems are not equipped with these kinds of functions which respond to the acoustic properties in the usage environment and automatically apply equalization.

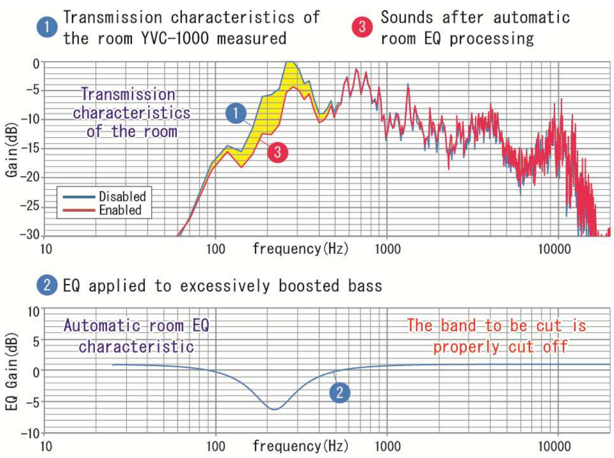
< Performance validation >

Validation results:

1) Case in an excessive bass boost environment

The difference in transmission characteristics between the presence and absence of automatic room EQ was validated in an excessive bass boost environment. The transmission characteristics of a room are the audio frequency response of sounds transmitted from the speakers to a user location that are determined by the shape of the room and the positional relationship between the speakers (YVC-1000 built-in and external speakers). This validation data was measured via the premise that the user's location is equal to the microphone's location. Graph 2-6 expresses that the higher up the longitudinal axis the frequency band goes the more that the loud sounds are conveyed.

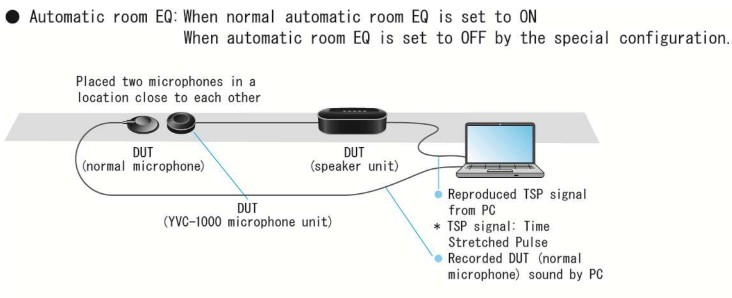
(Graph 2-6)



As the graph shows in the shaded in yellow section between ① and ③, it is clear that excessively-boosted bass is properly suppressed by the automatic room EQ. At these times, overall gain adjustments are performed so that the volume sensation itself does not change.

Data measuring method:
In the configuration as in Figure 2-8, voices before and after processing with the automatic room EQ were measured by turning it ON and OFF via the software.
* The above validation results were obtained using YVC-1000 firmware Ver.1.05.

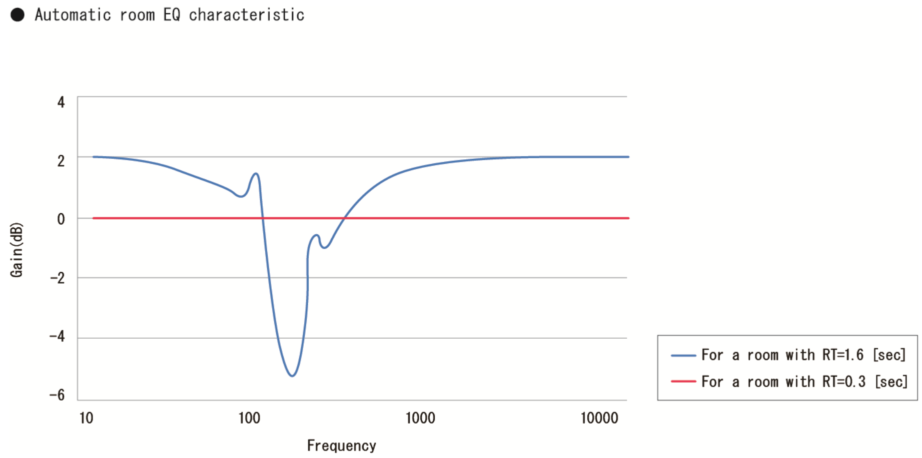
(Figure 2-8)



2) For large reverberation environments

The automatic room EQ property of cutting off a band subject to large reverberations was validated.

(Graph 2-7)



Graph 2-7 expresses the automatic room EQ property differentials when the automatic room EQ is operated in rooms with virtually no reverberations (100 ms reverberation time), and when it is operated in rooms with many reverberations (1100 ms or greater reverberation time). In this manner the YVC-1000 automatic room EQ automatically evaluates the reverberation effects in rooms, and in response to them it attenuates the great frequency bands of the reverberation effects to eliminate “muffled and muttered voices,” to actualize clear voice sound pickups.

Chapter 3 Two Unique Technologies to Enhance Six Sound Processing Capabilities

1. Human Voice Activity Detection (HVAD)

1.1 Outline and purpose

HVAD is a technology which discerns whether or not human voices are included in voice signals that are picked up by YVC-1000.

1.2 Operations and mechanisms

HVAD operates in the basic signal processing technologies in the three following functions to dramatically enhance accuracy.

- Automatic tracking
- Noise reduction
- Automatic gain control

This section describes the HVAD mechanisms in combination with the above signal processing technologies using the signal processing flow charts.

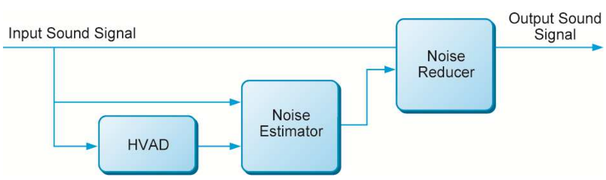
* For details about individual signal processing technologies, refer to the corresponding pages in Chapter 2.

1) In combination with automatic tracking

YVC-1000 can pick up voices clearly even in an environment where voice locations change among participants. An audio source location is estimated by the microphones array control device that consists of three microphone elements. When the sound source orientations are presumed, HVAD determines whether those sound sources are human voices or not and uses those results to capture the orientations of the isolated sounds and steady noises as the speakers' locations to dramatically reduce mistaken identifications.

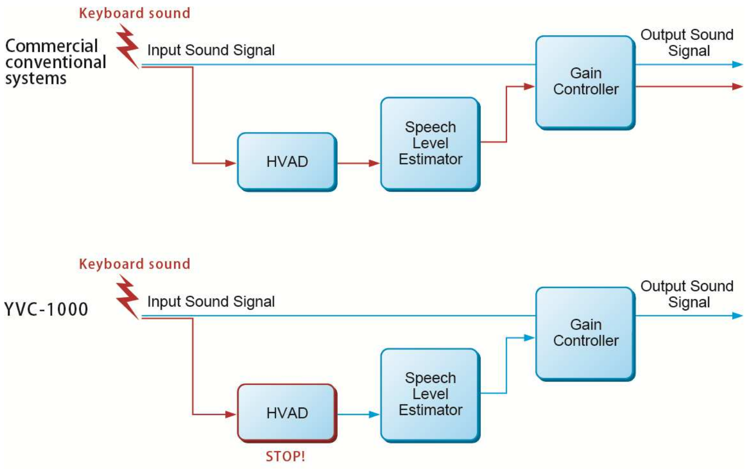
2) In combination with noise reduction

The noise reduction function is required to estimate noise components in order to eliminate only the noise components from voices. The method to estimate a steady signal as a noise component is commonly used, but some audio signals (such as prolonged voice "Aaa..." in conversation and music) may be eliminated as misidentified noise components. Since YVC-1000's noise reduction can distinguish human voice components from noise components using the HVAD determination results, noise reduction performance is improved.



3) In combination with automatic gain control

The automatic gain control function properly corrects the sound level of voices picked up by a microphone. This requires the function to accurately pick up the volume of human voices before they are corrected. YVC-1000 can estimate the level of human voices via HVAD with high precision by distinguishing between signals containing human voice and signals consisting of only noises. The result is that auto gain control performance can be stabilized.



Since human voices cannot be distinguished if there is no HVAD, the auto gain control might respond to abrupt or steady noises. For example, if this function responds to intermittent low sounds (e.g. from keyboard typing) while nobody is saying anything, then the sound level will be underestimated (recognized as a low voice), thereby significantly increasing the microphone gain. As the result, there is a possibility of sending objectionable sounds (intermittent noise) other than voices to the other communicating parties. YVC-1000's automatic gain control minimizes such a risk thanks to the presence of HVAD.

2. Automatic Audio Tuning Function

2.1 Overview and purpose

This function learns the acoustic environment inside the room, and automatically optimizes the parameter settings of the signal processing technologies, etc. (adaptive echo canceller/automatic room EQ).

● Conferences can be started in conditions with optimal settings

In general, the learning by the echo canceller and the noise reduction, etc. would progress after a certain extent of time had passed whereas conversation with the distant party had taken place, then the advanced signal processing capability could be demonstrated. However if this function is used, then conferences can be conducted in conditions with optimal settings immediately after they are begun. When the automatic audio tuning function is implemented, all of the YVC-1000's acoustic settings are optimized via the reproduction of the measuring sounds.

● Adjustment with an external speaker connected

This function is particularly useful when connecting an external speaker since the difference in delay between the external and built-in speakers and frequency response can be corrected. (Refer to 3) to 4) of 2.2 in Chapter 3.)

If YVC-1000 detects acoustic problems during normal operations, or during the automatic audio tuning processes, the users are notified of the anomalies and their causes via the lighting up and blinking of the tuning fork's button, and the audio guidance. With conventional conference microphone and speaker systems, what was causing the voice degradations as in the following example could not be picked out, thus users had difficulty in dealing with those causes. However since YVC-1000 specifies those causes and notifies the users of them, tiptop conditions can be brought to conferences.

Example of audio guidance

Meaning	Tuning fork view
A noise source exists near microphone	Lights up orange
Microphone and speaker are placed too close	Lights up orange
Difference in delay between built-in and external speakers is too large	Blinks orange light at high speed

2.2 Operations and mechanisms

The automatic audio tuning function optimizes the following four function parameters by learning about the acoustic environment in a room where it works.

1) Adaptive echo canceller
2) Automatic room EQ
3) Correction of the difference in delay between built-in and external speakers
4) Correction for the frequency response of external speaker(s)

1) Adaptive echo canceller

YVC-1000's echo canceller is a filter-type adaptive echo canceller that learns about the acoustic characteristics between the speakers and microphones. Since this is different than a suppressor type echo canceller, both parties can simultaneously hold pleasant conversations. When the automatic audio tuning functions are implemented, the measuring sounds are reproduced from the speakers, and because the adaptive filter can sufficiently learn by this, this can bring to conferences a simultaneous conversation efficiency by both parties which is optimized immediately after conversing has started.

2) Automatic room EQ

In normal operation, automatic room EQ calculates the optimal parameter settings for approximately several tens of seconds to several minutes while gradually adapting to the operating environment. Upon execution of the automatic audio tuning function, the measuring sounds are reproduced from the speakers to immediately update the optimal settings of automatic room EQ parameters. This allows YVC-1000 to reproduce voices with the best audio quality from the start of conferences.

3) Correction of the difference in delay between the built-in and external speakers

If any delay between the input and output of external speakers exists (e.g. when using a TV as an external speaker), it may cause a time lag between the voices being simultaneously reproduced from the external and built-in speakers. The delay times of external speakers are measured by executing the automatic audio tuning function.

YVC-1000 sets an appropriate delay in sound from the built-in speakers in accordance with the measurement results to correct the differences in delay between the external and built-in speakers. This setting value is retained even if the power is turned off.

4) Correction for the frequency response of external speaker(s)

If the frequency responses are greatly different between the external and built-in speakers, the localization of sounds simultaneously reproduced from the external and built-in speakers will be scattered. For example, it is conceivable that there are situations that would be illusory whereas it would seem that the mid-low range sounds are being emitted from the external speakers, and that the high range sounds are being emitted from the internal speakers. As a countermeasure for this type of phenomenon, YVC-1000 adaptively corrects the frequency properties in regard to the signals being sent to the external speakers, thus it can cause the frequency properties of the external speakers to get near to the frequency properties of the internal speakers. The frequency response characteristics of the external and built-in speakers are measured by executing the automatic audio tuning function and reproducing the measuring sounds. This setting value is retained even if the power is turned off.

Conclusion

In chapter 2, the findings which validated the performance of each of the YVC-1000's signal processing technologies, confirmed excellent performance in all of the acoustic elements that influence the actualization of smooth remote conference communications.

Different than products which have targeted individuals or a few people, this product was envisioned for group usage in mid- to large-sized conference rooms whereas performance compatible with more advanced usage environments requires sound pick up performance and accurate conversant focus, and noise countermeasures, etc. YVC-1000 was envisioned and designed with attention to these details, furthermore optimal sound quality can be sustained via the simple UI design which does not demand users to configure complicated settings. This product is suited for supporting stress-free and smooth communications in the various scenarios of web conferences that are going to accelerate in the future.

Afterword

Yamaha is assuming the important role of urging market expansion in web conferences which are going to become the mainstream in the future via the high quality sound technologies indicated in this white paper, thus Yamaha thinks that it can actualize smoother and more fulfilling communications. Yamaha is aiming for remote communications that do not cause distances to be felt via the true-to-life realistic feeling that any conversing taking place is just as if it were in the very same room, thus hereafter we shall strive towards developing technologies.

YAMAHA CORPORATION

10-1, Nakazawa-cho, Naka-ku, Hamamatsu City, Shizuoka Prefecture, 430-8650

http://jp.yamaha.com/products/communication/