The Basics of High Fidelity
Part 8: Telephony

In the previous seven parts we have deepened our understanding of HiFi and come to the conclusion that HiFi stands for "naturalness" in the sense that we are striving for transparency between recording and reproduction. We have seen two different ideals we can strive for, the "Here and Now" (Augmented Reality) or the "There and Then" (Virtual Reality). The "International Telecommunication Union" (ITU) has defined a clear ideal for telephone links, we should strive for a link that represents the situation of two subjects communicating at a distance of about 1 meter in a silent, low reverberant, room. This represents something between the "Here and Now" and "There and Then" as we have discussed in our previous seven papers. One would expect that there are no real technical problems in achieving this ideal. We can take a high quality microphone and loudspeaker, couple them closely to our mouth and ear, and we should be able to come close to the ideal telephone link.

Unfortunately there are three snags. The first one is the microphone coupling, it should be as close as possible. In a classical handset the distance between the microphone and mouth is less than 1 cm. The reason why this distance should be as close as possible is the fact that as soon as the distance increases room reverberations and background noise start to degrade the speech signal. Combine this with the fact that many rooms are of low quality (high levels of reverberation) and often have high levels of background noise and you will understand why the use of modern communication apps such as Skype, Facetime, Zoom, or whatever modern app we install on our computer, have such a low speech quality. They all suffer from a too large distance from the microphone as a result of using a camera, leading to a hollow sounding voice and clear audible background noises. Also the fact that with the introduction of mobile telephony people have started to use their cell phones in the harsh acoustic environments, even in bathrooms, further adds to the close mic coupling problem. Classical handsets are designed with close coupling of microphone and loudspeaker to our mouth and ear, while modern smartphones use a loose coupling, leading to poor quality recording and play back. This failure of providing a close acoustic coupling is then compensated by advanced signal processing that in most cases will introduce new types of degradations in the speech signal. You may ask yourself why the hollow sound degradation, as perceived with larger mic distances, is so much more disturbing than when I place my ears at the same place as the microphone of the computer. The answer is binaural decolorization, a process that suppresses the reverberation of the room we are sitting in. So if you are using these types of app's place a microphone close to your mouth at the same position as the microphone of a classical handset.

The second snag lies in the fact that telephony is not only a question of passive listening, like in HiFi, but also of active talking. And people talking love to hear their own voice in high quality [1]. In the case of the live ITU reference condition this is no problem, we hear our own voice in a natural way. A telephone link however introduces a major problem in talking, my voice is picked up at the other side of the connection and transported back to my own ear sometimes resulting in a disturbing echo. You may be familiar with it if you make long distance calls, where the network delay is so large that echo needs to be suppressed, or even better cancelled. In modern packet switched networks, either fixed or mobile links, delay is always so large that echo cancelation is mandatory. In addition when you make a phone call in a noisy environment you will

automatically press the phone against your ear, blocking the acoustic feedback path of your own voice. In classical handset design this blocking is compensated by a direct path from the microphone to the loudspeaker, the side tone path, simulating the natural acoustic feedback path from my mouth to my ear.

The third snag lies in the fact that telephony in a highly interactive scenario requires a low end-to-end delay (one way preferably below 200 ms). For long delays, above 400 ms, users will have severe difficulty in interrupting each other in a natural manner and they have to adapt their interaction strategy. You can check the delay by interactive counting, you start with 1 as you press start on your stop watch, your partner answers 2, you continue with 3, ….etc. until you stop at 10. This takes about 4.5 seconds in a live situation and any second extra is 100 ms of one way delay on the connection.

So we see that making a high quality telephone link is not as simple as we may think. Especially when using Skype, Facetime, Zoom, …., which all suffer from a too large distance from the microphone, we get unacceptable low conversational quality. In modern telephone like applications all mentioned degradations are suppressed by smart signal processing. Unfortunately each problem we try to solve introduces a new one. An example of such a disturbing new degradation can be identified as “back ground noise switching”. Each time we start to talk we can hear a clear change in the noise we hear in the connection. Another example is double talk degradation, when both partners talk one of the voices is suppressed. And a solution towards perfect noise suppression is counterproductive, users will start to complain that they are unsure as to whether the connection is lost.

So what should we demand from a high quality telephone connection? We must take into account the three main aspects of the conversational quality:

1) Listening quality (passive), how how do I perceive the voice from the other side of the link (noise, distortion).
2) Talking quality (active), how do I perceive my own voice in clean and background noise situations [1] (echo, side tone, background noise switching).
3) Interaction quality (active), how well can both parties interact with each other. It is composed of two contributing factors, delay and double talk distortions.

With modern coding techniques and close coupling microphones the listening quality can be of high quality. Modern cell phones provide a so called HD (High Definition), or even SHD (Super High definition) voice feature, which extends the audio bandwidth from 3.5 kHz (Standard Definition) to 7 or even 14 kHz. One should however realize that a human voice has little energy below 150 Hz and above 6 kHz [2], so background noise may become more of problem than with SD (Standard Definition) voice. Providing high talking and interaction quality is more complicated, especially in combination with (S)HD voice. A HiFi Phone is still not available today.

[1] S. R. Appel and J. G. Beerends, “On the Quality of Hearing One’s Own Voice,” J. Audio Eng. Soc., vol. 50, pp. 237-248 (2002 April).

[2] H. Fletcher and R. H. Halt, “The perception of speech and its relation to telephony,” J. Acoust. Soc. Am, Vol. 22, No. 2, march 1950.

John G. Beerends
Published in Hifi Video Test 8/2008 (in Dutch), translated and updated over the period 2012-2020.